NASA/CR—1998-207408

Technology Directions for the 21st Century

Volume IV

Giles Crimi, Henry Verheggen, Robert Botta, Heywood Paul, and Xuyen Vuong
Science Applications International Corporation, McLean, Virginia

Prepared under Contract NAS3–26565

National Aeronautics and
Space Administration

Lewis Research Center

May 1998

PREFACE

In this rapidly changing world, effective governmental organizations must constantly anticipate and regularly plan for the future. For NASA, the process of doing this has become particularly challenging. Global economic strategies have greatly increased the need for cooperative partnerships to provide adequate resources for space science and exploration programs. Today's technological advances provide ample opportunities to improve the ways in which NASA carries out its responsibilities and serves its customers. NASA's future technology emphasis must focus on dual-use activities; that is, technology must be transferred or leveraged with other government agencies, industrial partners, international counterparts, and academia to share costs, exchange technical knowledge, support a growing economy, and provide technological leadership for our Nation. The competitiveness of commercial ventures, such as those ignited by the deregulation of the telephone industry, has shifted the burden of continuously developing cutting-edge technology from government to the commercial sector.

This is the fourth volume of a series of technology trends and applications reports. These reports focus on the continuing effort to predict and understand technology trends to better plan research and development strategies. The objectives of this series of documents are to: (1) validate, revise or expand relevant technology forecasts; (2) develop applications strategies to incorporate new technologies into our programs; and, (3) accomplish Agency goals while stimulating and encouraging development of commercial technology sources. In fact, this report provides a focal point for the previous three volumes of Technology Directions for the 21st Century . The first three chapters of this volume continue the theme of identifying the trends of some selected advanced technologies. The final chapter, however, identifies how advanced technologies have critical application to future satellite communications systems. Here the bridge is completed from evaluating technology trends for improved processing performance, storage and optical techniques to specific systems' applications of advanced technologies. The data and information herein are current as of December 1997.

For terrestrial networks, new technologies continue to support the ever increasing demand for bandwidth. In addition, a shift from private to public networks is underway as public networks achieve higher levels of availability and reliability, and offer a flexible menu of services. As for satellite communications, there are today over $18B in investments for low Earth orbiting data and voice communications systems. In general, the distinctions among long- and short-haul carriers, telephone and cable operators, wireless and wired networks, data and video are becoming increasingly blurred. Private investments in communications infrastructure now significantly exceeds the budget for communications investments in agencies the size of NASA. With a market of over three billion people on this planet without current access to a telephone, revenues for satellite services worldwide are expected to grow to more than $37 billion per year by the year 2000. There are currently 55 distinct plans identified from North America, Europe, Asia, and Africa involving the launch of nearly 1,300 Ka band satellite networks for the next generation of communications satellite systems.

It is clearly incumbent upon NASA to position the Agency to take full advantage of available technology and to use trends and projection data for the purposes of planning, building, operating and sustaining information handling systems. To maintain technological leadership, we must know where technology is going, where our technology stands with respect to the rest of the world, and how best to capitalize on areas of maximum leverage with the investments of our commercial and academic partners. It is within this context that this report has been prepared, as part of an on-going effort to plan for the future of space communications and information systems.

Comments from interested parties would be especially appreciated, and may be directed to Dr. Albert R. Miller, Office of the Deputy Under Secretary of Defense (Space Integration), at (703) 325-3279, or to Ms. Denise S. Ponchak, NASA Lewis Research Center, at (216) 433-3465.

Denise S. Ponchak

Denise S. Ponchak
Chief, Project Development and Integration
Space Communications Office
NASA, Lewis Research Center

ACKNOWLEDGMENT

This work was performed by Science Applications International Corporation (SAIC), Software and Systems Integration Group, in McLean, Virginia for the National Aeronautics and Space Administration (NASA). This report summarizes a series of investigations of various technologies and systems which are important to NASA's long range planning for operational communications systems and capabilities.

NASA is pleased to acknowledge the following SAIC personnel who contributed to this effort: Dr. Giles Crimi (Program Manager); for Chapter 1, Mr. Henry Verheggen (Principal Investigator); for Chapters 2 and 3, Dr. Robert Botta (Principal Investigator); for Chapter 4, Messers. Henry Verheggen and Heywood Paul, and Dr. Xuyen Vuong (Principal Investigators).

The technical direction for this work was provided by Ms. Denise S. Ponchak, NASA Lewis Research Center, and Dr. Albert R. Miller, Office of the Deputy Under Secretary of Defense (Space Integration).

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1. DATA COMPRESSION TECHNOLOGY TRENDS

SUMMARY

Data compression is an important tool for reducing the bandwidth of communications systems, and thus for reducing the size, weight, and power of spacecraft systems. For data requiring lossless transmission, including most science data from spacecraft sensors, small compression factors of two to three may be expected. Little improvement can be expected over time. For data that is suitable for lossy compression, such as video data streams, much higher compression factors can be expected, such as 100 or more. More progress can be expected in this branch of the field, since there is more hidden redundancy and many more ways to exploit that redundancy.

1. INTRODUCTION

Data compression is the encoding of information (e.g., text, images, video) into as few bits as possible for transmission, together with a decoding mechanism that permits reconstruction of the original data with acceptable quality. Data compression is a standard tool for the reduction of bandwidth in most modern communications systems and can have a major impact on the size, weight, and power of a spacecraft communication system. Compression techniques are often distinguished as being either lossless or lossy. Lossless techniques allow the exact reconstruction of the original data after it has been compressed. This may be important for many types of scientific data for which the redundancy characteristics are unknown in advance. Lossy compression allows the loss of some information in exchange for potentially higher compression factors. For example, most compression techniques for video are lossy, but attain much higher compression factors than would a lossless technique. This report presents an overview of modern data compression techniques across the spectrum of data types, and discusses what sort of progress and innovations may be expected in this field.

2. LONG RANGE TECHNOLOGY TRENDS

The main approaches to data compression have been known for several decades and progress in much of this field is gradual and incremental. Data compression factors have improved over time at different rates, depending on the type of data being compressed. Low-redundancy data cannot be compressed much, by definition, whereas high-redundancy data can be compressed to a greater degree. Historically, it has taken longer for researchers to find ways to exploit more of the redundancy in high-redundancy data such as imagery and video. Further progress in this area can be expected, while little progress in the area of low-redundancy data can be expected.

2.1 ANALYSIS OF TRENDS

2.1.1 Text Or Symbol Compression

Character-oriented [HELD96]

Character-oriented compression has been used for many years for the compression of one-dimensional streams of symbols, such as text files. It uses techniques such as bit mapping, run-length encoding, and byte-pair encoding. These techniques basically replace redundant character patterns with shorter character codes. Run-length encoding, for example, is an important technique that is used in many compression schemes. It is based on the concept of replacing repeating characters with a code for the number of repetitions, the run-length. The various character-oriented techniques, when applied to computer files typically result in compression factors of about two.

Statistical [HELD96, LYNC85]

Statistical compression techniques are based on an analysis of the probability distributions of symbols to allow the use of optimal length codes. The model for this family of techniques is Huffman coding. A Huffman code requires prior knowledge of the probabilities of occurrence of the symbols. Given such knowledge, the code is optimal in that it results in the shortest average code length of any statistical technique. The Huffman code is constructed such that more frequently occurring symbols are assigned shorter codes. Huffman codes must be different for different content types. For example, English text has a different alphabetic distribution than C source code. For information with unknown characteristics, adaptive Huffman coding techniques have been developed. These examine a large initial block of data, do a statistical analysis, and construct the code in real time. The decoding table is then sent to the other end of the communications link ahead of the compressed data. The table is then updated on an incremental basis, as the statistics of the symbol stream change.

Related compression techniques are Shannon-Fano coding, arithmetic coding, and Microcom Networking Protocol (MNP) compression. MNP is a de-facto standard used in millions of modems. MNP 5 compression uses a combination of run-length encoding and an adaptive statistical encoding technique, and achieves compression factors between 1.5 and 2. MNP 7 is an improved technique that uses a Markov model to predict the probability of the next occurrence based on the previous character, and performs adaptive Huffman coding. It uses run-length coding to compress duplicate character sequences. MNP 7 offers an improvement of 15-25 percent over MNP 5.

Dictionary-based String [HELD96]

Most dictionary-based string compression mechanisms trace their roots to Ziv and Lempel's algorithms published in papers in 1977 and 1978, known respectively as LZ77 and LZ78. The method became widespread after Terry Welch published a modification in 1984, which is known as LZW. LZW became the basis of common applications such as the V.42bis modem compression (BTLZ), Compuserve's Graphics Interchange Format (GIF) image file format compression, and commercial disk compression programs. These techniques replace strings of symbols rather than single symbols with codes for the strings. This can be shown to approach entropy (the level of zero information redundancy) more closely than single character encoding. A dictionary of strings is constructed, and the strings in the data stream are replaced with a code for the position of the string in the dictionary. The dictionary is constructed on a dynamic basis, that is, the strings are "learned" and added to the dictionary as more data is processed. Once the dictionary is filled, a "pruning" algorithm is applied to maintain a constant size for the dictionary.

2.1.2 Voice Compression [LYNC85, ASPI96, BELL82]

Voice compression techniques are a major sub-discipline of compression technology. The driving economic force in this field has been voice signal compression, as might be guessed in view of the amount of communications traffic of this type. The simplest audio compression technique is delta-encoding, which is the transmission of only the difference between signal sample amplitudes. It is an example of the class of predictive coding systems, which includes all the techniques discussed here. In predictive coding, a predicted value for the next sample is subtracted from the actual value to get an error. The error is then encoded so as to achieve compression. A commonly used extension of delta-encoding is continuously-variable slope delta (CVSD) encoding, in which the increment of the delta is adaptively adjusted to the waveform. This system achieves a data rate of 16 kilobits per second for a 3 kHz voice channel, but at medium fidelity [ASPI96].

Pulse-code modulation (PCM) is universally used for digitized telephony signals. A standard digital voice telephony channel (e.g., International Telecommunications Union (ITU) G.711) consists of a PCM bit stream. The voice signal is sampled at a rate of 8 kilosamples per second, and the dynamic range is logarithmically compressed to 8 bits of amplitude, resulting in a bit stream of 64 kilobits per second. A more advanced technique is adaptive differential pulse-code modulation (ADPCM). ITU standard G.726 specifies an ADPCM system with optional rates of 16, 24, 32, and 40 kilobits per second [ASPI96]. Some digital solid-state voice storage devices use ADPCM, for example the Dallas Semiconductor DS2270.

Higher compression factors can be achieved by taking advantages of the specific characteristics of the human vocal tract and ear. The highest degree of compression of voice signals has historically been achieved by linear predictive coding (LPC). In LPC systems, the human vocal tract is modeled as a discrete time-varying linear filter. The coefficients of the linear filter are transmitted instead of the waveform samples. Early military applications of LPC achieved 2.4 kilobits per second data rates compared to 64 kilobits per second for an uncompressed 3 kHz voice channel, but with low fidelity. Such low data rates were important for digital encryption and digital communication, since processors and modems required low data rates at that time – the 1970s and 1980s. The drawback was that a significant amount of speaker recognition, important in military communications, was lost.

More recent efforts have focused on improving the fidelity of LPC for commercial applications, such as digital cellular telephony. Earlier LPC encoders used white noise and pulse trains to excite the filter. More recently, other methods of excitation have been found to produce higher fidelity. The ITU has selected a number of these advanced LPC systems for standardization. For example G.728 and G.729 specify two 16 kilobit-per-second code-excited linear prediction (CELP) systems. The code referred to here is a selection from a codebook of predefined prototype signal vectors. The Department of Defense recently selected an algorithm for the new 2.4 kilobit-per-second federal standard. It is called the Mixed Excitation Linear Predictive (MELP) Vocoder, developed by Atlanta Signal Processors, Inc. Mixed excitation refers to the use of multiple types of excitation waveforms [ASPI96].

Compression of voice signals can be pushed significantly lower than 2400 bits per second. Vector quantization LPC techniques have achieved voice encoding at rates as low as 800 bits per second [LYNC85].

2.1.3 Facsimile Compression [HELD96]

Facsimile essentially consists of a 1-bit gray-scale image of 1100 scan lines of 1730 pixels, or 1.9 megabits. The large amount of data needed to represent a single page dictates the use of compression. Modern facsimile is based on the ITU Group 3 protocol, a digital protocol compatible with standard modems. Group 3 compression uses a Huffman scheme to encode one-dimensional run-lengths of the same bit value. Rather than use an adaptive encoding technique, Group 3 uses codes that were derived from a set of standard documents defined by the ITU. They can thus be stored in lookup tables, reducing the amount of processing required, and hence the cost of the fax machine. Since there is a high degree of correlation from scan line to scan line, Group 3 allows the use of a two-dimensional coding scheme to take advantage of this, called modified read coding.

Facsimile is the most common form of image communication today. The popularity of fax illustrates the economics of compression. An improvement of 10 percent in the compression factor does not seem like much, but if applied across all fax machines worldwide, the aggregate savings would be great. Historically, the improvement in transmission times for a single page has been significant, from minutes to seconds, and this in turn has been a major factor in the proliferation of fax.

An obvious step for improving fax would be to incorporate character recognition, in which text images would be converted to character codes. Such a system would have to be a hybrid, in which recognized characters are encoded as characters, but parts of the page where non-text images are present would be transmitted as standard fax.

2.1.4 Still Image Compression

Compression of still imagery is a dynamic field of research and development. There are two major categories of image compression, lossy and lossless. Lossless compression refers to algorithms that allow the exact reconstruction of the uncompressed image down to the pixel level. The use of lossless compression may be mandated in applications such as medical radiology, earth resources imaging, or reconnaissance imaging, where loss of information cannot be tolerated. Lossless techniques typically achieve a compression factor of about 2. For example, Dr. Pen-Shu Yeh of the National Aeronautics and Space Administration (NASA) Goddard Space Flight Center developed a lossless compression technique that is used extensively in NASA and the government and was first flown on the Mars Observer. It has a compression factor of 2 to 3. Note that this system is not specific to image compression [NASA96]. A commonly-used lossless image compression algorithm is the GIF. This uses a type of LZW compression and is therefore lossless [HELD96].

Higher compression is achievable when some loss of information is acceptable. A commonly-used lossy image compression algorithm is that defined by the Joint Photographic Experts Group (JPEG) of the International Standards Organization (ISO), and which is known simply as JPEG. It is for compression of monochrome or full-color, grey scale, digital still images. Options within the standard allow for the use of: lossy, discrete cosine transform (DCT); lossless predictive algorithms; or Huffman or arithmetic coding.

The baseline version uses the lossy DCT algorithm with Huffman encoding. Typical compression for the lossless algorithm is a factor of 2. For the lossy algorithm, a factor of 10-20 results in minimal visual degradation (visually lossless). Even higher compression in the 30-50 range can be selected, but with some loss of quality.

There are four modes of operation, sequential DCT-based, progressive DCT-based, lossless and hierarchical, but applications do not need to implement all these modes. In hierarchical mode images are encoded as a sequence of frames. Except for the first frame in the hierarchy all other frames can be difference frames that identify the differences between a child frame and its parent.

2.1.5 Video Compression

The Motion Picture Experts Group (MPEG) of the ISO has defined two video standards, MPEG1 and MPEG2, which are documented in standards ISO/IEC 11172 and 13818, respectively. MPEG1 was defined for VHS-quality applications, especially compact disk, read-only memory (CD-ROM), at 1.5 Mbps. It is based on Huffman encoding of video luminance and chrominance DCT frequency components. Humans derive more information from the luminance signal than the chrominance, so luminance is given preference in the quantization and in the Huffman code assignments. Also, most of the information is in the lower frequency components, so these are also given shorter Huffman codes. Further compression is achieved by processing interframe redundancies using predictive coding. Blocks of pixels in one frame can be predicted from blocks of pixels in a previous frame. Similarly, a motion-compression scheme is used, in which a search is conducted for blocks of pixels that have been translated to a new position. If a match is found, a motion vector for the block is encoded, rather than the whole block.

MPEG2 is a more advanced, high-compression system that is downwardly compatible with MPEG1, and also uses Huffman encoding of DCT components as its basis at the frame level. It is designed to transmit broadcast-quality video at 6 Mbps, while being optionally scaleable to lower quality levels at 4, 3, 2, and 1.5 Mbps. Interframe compression is also used. At still higher layers, MPEG2 offers a set of "profiles", that represent different degrees of temporal and spatial scalability, and allow mixing and matching to different bit stream rates, video formats, and content types.

The scalability features provide great flexibility. For example, multiple frame formats and sizes are supported (spatial scalability). As well, the protocol is divided into a low-resolution part and a high-resolution part. This allows a high definition television (HDTV) signal to be decoded by NTSC, PAL, and SECAM decoders. Temporal scalability is provided in that one signal can be displayed at multiple frame rates. Signal-to-noise scalability allows one signal to be compatible with multiple levels of quality. There is also a two-channel priority scheme, which allows a transport system to drop the lower-priority channel under congested network conditions, resulting in a reduced-quality image, but not a total loss of data. This is significant for asynchronous transfer mode (ATM) networks, in which cells may be discarded or lost.

MPEG2 also defines two types of transport, called the Program Stream and the Transport Stream. The Program Stream is for low-error-rate environments such as computer multi-media and digital storage. This allows multiple sources of audio, video, and data to be multiplexed within a single set of variable-length packets. The Transport Stream is for digital television. It allows multiplexing of multiple programs and an electronic program guide in a stream, using fixed-length, 188-byte packets [RUIU96].

The most recent video teleconferencing standard, H.324, includes a compression standard, H.263, for compression of video to 20 kbps [GOLD96].

2.2 FUTURE APPLICATIONS

2.2.1 Internet And Digital Video

The major economic force driving compression today is the explosion in demand for bandwidth on the Internet and the various schemes for digital video and digital multimedia distribution. Currently, an abundance of new products is becoming available for digital video based on MPEG2 and H.261. Many telerobotics applications could make use of these standards. For moderate quality, H.261 is a good choice, while MPEG2 works for higher-quality video.

2.2.2 Scientific Imaging

Not much improvement is to be expected in the field of lossless image compression, for the simple reason that not much redundancy is normally available for lossless removal. Application-specific satellite imagery such as military surveillance or weather imagery presents opportunities for compression that more generalized missions do not. As processors become more powerful per size, weight, and power, on-board processing could allow images to be reduced to aggregate measurements without the need to send the entire image. For example, if one is looking for man-made objects in a surveillance image, and one can define signatures for man-made objects, most of the background could be filtered out. If an earth-resources image can be segmented into statistical clusters based on terrain spectral reflectance, only the segment boundaries and their statistics need to be transmitted.

2.2.3 Very Low Bandwidth Compression

Very low bandwidth compression refers to techniques such as transmitting only the outline of a face to convey the essential facial information over a teleconferencing link. Another example is 800 bit-per-second voice encoding, which conveys the content information but not necessarily much of the speaker's distinctive voice characteristics.

An example of early very low bandwidth videoconferencing research involved finding an edge filter that extracted an outline facial image that retained the facial characteristics important to a human observer. The pseudo-Laplacian filter was found to preserve the most natural appearance, while converting an 8-bit grey-scale image to a 1-bit outline image, and achieving a compression factor of 20 [PEAR85]. Presumably, even greater compression could be achieved in a video stream of these images, by using inter-frame compression.

At present, MPEG4 is the standard that will address low bit rate video. MPEG4 is still in the process of being defined, but some key themes have been settled. It will cover an expanded scope of motion imagery and sound to include ways to combine computer-generated graphical and sound objects with conventional video imagery and audio. This "hybrid coding" offers improved compression opportunities, since object equations may be sent instead of images of objects. It will allow the use of model-based compression, in which a 3-D or 2-D model of the objects in a scene are estimated from a 2-D image or sequence of images. A recent paper discussed results of a model-based encoding scheme for human faces that used a 3-D and a 2-D model, and achieved a bit rate of 5 kbps for a 10 Hz frame rate [AIZA95].

One can imagine future video teleconferencing hybrid coding incorporating virtual reality, in which the background is synthesized as 3-D graphics with live human figures spliced into the virtual reality background.

2.2.4 Facial Recognition For Security

In the last few years, facial image recognition systems have emerged as working products for the security market. One successful method of face recognition called "eigenfaces" is based on modeling an individual human face as a set of differences from an ideal or prototype face model. A byproduct of this technique is that the facial images can be dramatically compressed, to less than two hundred bytes. Very large databases of facial images can be searched rapidly using this method. It also has application to video teleconferencing. This algorithm is available as a product called Sherlock marketed by Facia Reco, Inc. This technique will be discussed further below under the heading "Principal Component Analysis".

2.3 EMERGING TECHNOLOGIES

2.3.1 Wavelet Compression

A wavelet transform is a transform that uses localized basis functions rather than the cosines and sines of the Fourier transform, which are infinite in extent. Fourier-transform based compression operates on the fact that in the frequency domain, the information is sparsely distributed among frequencies for most images, so relatively few components are needed to encode the information. Wavelet transforms of images are even more sparsely distributed, because most images contain localized features, not periodic features. Therefore wavelet transforms can achieve greater compression. Wavelet transforms also do not exhibit the blocking artifacts of the DCT and other block-oriented approaches [BRUC96].

The Houston Advanced Research Center (HARC) is currently publicizing a wavelet transform called HARC-C which is capable of compression up to a factor of 200 on a single image. Ball Aerospace has adopted this algorithm for use in its space image sensor business [HARC96].

2.3.2 Fractal Compression

Fractals are functions that exhibit self-similarity over a range of scales. For example, a historical graph of the Dow Jones Industrial Average exhibits random fluctuations that look similar, whether it is examined over a period of months, days, or minutes, and thus exhibits a fractal shape. Wavelets can be fractal, so there is a relationship between fractal and wavelet compression. Natural objects have a great deal of fractalness to their shapes, so images of natural scenes can theoretically be described as a superposition of fractal shapes. Given a sufficiently large library of fractal functions, one can decompose an image into a set of equations describing the fractal functions, resulting in very large compression factors.

Iterated Systems, Inc. has patents on fractal compression, and has the only commercial products that perform fractal compression. The technique is advertised as capable of achieving compression factors between 100 and 1000. Factors of 100 seem to be routinely achievable. Fractal compression does not show blocking artifacts, and is scale-independent. Scale-independence means that the decompressed image has the same quality over a range of magnifications. Fractal compression imposes a high computational load on the encoding side, and has thus been typically limited to applications such as CD-ROM storage, in which the fractal transform is computed only once, but will be decoded millions of times [ITER96].

2.3.3 Model-based Compression

Model-based compression, discussed briefly above, is a completely different, but difficult to implement, approach to compression. It is based on the simple idea that an image can be described as a view of the set of 3-D objects rather than a superposition of random 2-D shapes and textures. Finding a model-based encoding entails two steps, first finding the objects in the scene, and second, modeling the objects with simple mathematical expressions. The general problem of finding objects in a scene is extremely difficult, related to many other hard problems in artificial intelligence (AI) and pattern recognition. The first experiments with model-based compression have typically involved simplifying assumptions. For example, in video teleconferencing, one could assume that the only object in the scene is a human face or head. Then, a general model of the head is constructed. An image is analyzed to find the head and its orientation. The individual characteristics of the person in the image are described as parameters of the generalized model. This results in a large compression factor. However, even larger compression occurs for subsequent images, in that they will be described as motions or distortions of the original model [AIZA95].

2.3.4 Principal Component Analysis

Very high compression factors can be obtained on human face images using a model-based technique that uses principal component analysis (PCA) (also called the Karhunen-Loeve Transform). This technique uses an information-theory approach to face recognition. It extracts the information content in a face by capturing the variation in a collection of face images, and then uses this information to encode and compare individual faces. The variation in the collection of face images can be quantified as a matrix in which each matrix position is the covariance between two pixel positions. If each pixel is thought of as a dimension in a multidimensional feature space, the covariance describes the distribution of measurements in that dimension. If the distribution is shaped like a Gaussian function in each dimension, then it has principal axes along which its shape is optimally specified. PCA is the operation performed on the matrix to find these principal axes. Once the operation is performed on a database of images, only a few of the features turn out to be important. This fact is exploited to achieve compression. A 16,384-byte image can be compressed to about 80 bytes – a factor of 200 [TURK91].

2.3.5 Compression, Computing And Artificial Intelligence

The use of a statistical pattern recognition technique such as PCA illustrates that there is a connection between compression and problems in artificial intelligence. This connection has been noted and proposed as a basis for a research program by Gerry Wolff at the University of Wales [WOLF96]. Wolff states that,

> "Information may be compressed by searching for redundant information and removing it wherever it is found. This means searching for patterns which match each other and merging or 'unifying' repeated instances of any pattern to make a one. Since there is normally an astronomically large number of alternative ways in which patterns may be matched and unified, it is necessary to use some kind of metrics-guided search ('hill climbing', 'beam search', etc.) or otherwise to restrict the search space in some way."

This leads to a general formulation of computing, AI, and compression as "pattern matching, unification, and search". Lossy compression implies that one is willing to lose some information but not the information "of interest". Automating the segmentation of information into that which is of interest and not of interest is essentially a problem of artificial intelligence.

Another interesting case in point is speech compression through speech recognition. A successful real-time speech recognition system can be thought of as system for compressing a 64 kbps data stream to a few bytes per second.

3. RESULTS

3.1 IMPACT OF COMPRESSION TECHNOLOGY ON NASA

Compression is a very important tool for communications for space applications. A lossless compression factor of two can have a major impact on the design of a spacecraft in terms of the size, weight, and power required to support a given link bandwidth. However, lossless compression by its very nature cannot be expected to improve significantly in the future. Lossy compression, especially for imagery, has improved dramatically in recent years. As stated above, the degree of lossy compression achievable depends on an ability to automatically sort out the information of interest. In this sense, further progress in lossy compression depends at least in part on the increasing sophistication of algorithms to recognize patterns, i.e., artificial intelligence. Model-based compression, in particular, relies heavily on AI techniques. The example of a speech recognizer, which converts a 64 kbps stream to a few bytes per second illustrates how potentially powerful AI can be. Clearly, there is a lot of room for progress in this direction.

In an era of flat or declining NASA budgets, there will be an emphasis on maximum use of commercial compression standards. At this time there is a large amount of private investment funding for lossy compression techniques for imagery and video. NASA can benefit from this work. NASA can also benefit from the compression standards that have been developed, and the commercial products based on them. While there is inevitably a lag between standards and the state of the art, there are many compatibility and interconnectivity benefits to using commercial standards.

The largest amount of data flowing over NASA communications links will be earth resources imagery. This imagery cannot usually be compressed by lossy techniques, because every pixel may contain useful scientific information. While this imagery may not have much redundancy on a local scale, there is redundancy on a global scale. Two images of the same patch of ground taken at different times will not be identical, but there may be enough commonality for differential encoding to provide significant compression. Undoubtedly, other global patterns such as seasonal changes could be exploited for compression.

While optical-fiber-based communication will eventually bring enormous bandwidth to the desktop and to the home, lessening the need for compression, in the radio-frequency domain, spectrum and bandwidth will always be relatively scarce. Although various techniques of spectrum re-use, and migration to higher frequency bands will continue to make additional spectrum available, compression will always be an important tool in RF.

3.2 TECHNOLOGY ROADMAP

The following is a technology roadmap, or hypothetical timeline, that summarizes projected developments in compression technology to the year 2020.

1996-2000: No major advances in lossless compression can be expected. In this period there will be widespread use of commercially-developed lossy video and image compression standards such as MPEG2 and H.261. Wavelet and model-based compression will emerge from the laboratory and be applied to real applications. Initial deployments of model-based compression will be in specialized applications such as video teleconferencing.

2000-2010: Artificial intelligence techniques for scene analysis, object segmentation, and model-based compression will be applied to more general image and video applications. A heavy use of computer-generated scenes can be expected in the video field. This artificial imagery will be mixed with real-world video, with improved compression as one of the benefits. Virtual reality teleconferencing may become commercially viable, although video teleconferencing to date has been a commercial disappointment.

The deployment of wideband digital communications service to the home will gain momentum in this period, and may affect the economics of compression. When fiber optics bandwidth becomes widely available, the need for compression for the mass market may plateau.

2010-2020: Developments in this time frame can only be speculated on. Artificial intelligence will presumably play an ever larger role in space applications. This means that spacecraft, planetary rovers, and other sensor platforms will be increasingly autonomous, and will need to rely less on communications of control information to earth. This is in essence a form of compression. On the other hand, the amount of imagery and other sensor telemetry sent from space platforms will probably continue to increase at a geometric rate. Bandwidth scarcity at radio frequencies may precipitate the use of optical free-space communications or on-board nanometer-scale storage.

Since traffic always expands to fill the available bandwidth, it is unlikely that the need to perform compression to the maximum extent will diminish. On the other hand, the benefits gained from further improvements in compression techniques will eventually plateau.

3.3 FEASIBILITY AND RISK

Data compression is a low-risk, high-payoff technology that trades size, weight, and power for software, and is becoming a standard feature of most communications protocols. Most compression needs can be met using off-the-shelf commercial standards and products. Specialized or NASA-unique applications may require the development of custom algorithms. There will always remain a need for public funding of advanced research in this field, and there is a potential dual payoff – both the commercial and government sectors can benefit. Visual communications is poised for an explosive growth in the next few years, especially within the commercial sector. Compression research will be a well-financed field of study.

References

[AIZA95] Aizawa, Kiyoharu and Thomas S. Huang, "Model-Based Image Coding: Advanced Video Coding Techniques for Very Low Bit-Rate Applications," *Proceedings of the IEEE*, February 1995, pp. 259-271.

[ASPI96] Atlanta Signal Processors, Inc., World Wide Web at http://www.aspi.com, 1996.

[BELL82] Bellamy, John C., "Digital Telephony", John Wiley & Sons, New York, 1982.

[BRUC96] Bruce, Andrew, et al., "Wavelet Analysis," *IEEE Spectrum*, October 1996, pp. 26-35.

[HARC96] from World Wide Web at http://www.harc.edu, 1996.

[HELD96] Held, Gilbert, "Data and Image Compression," John Wiley & Sons, Chicester, England, 1996.

[ITER96] from World Wide Web at http://www.iterated.com, 1996.

[GOLD96] Goldberg, Lee , "Advanced POTS Modems: Mr. Moore, Meet Mr.Shannon," *Electronic Design*, Oct. 1, 1996, pp. 77-88.

[LYNC85] Lynch, Thomas J., "Data Compression, Techniques & Applications," Van Nostrand, Rheinhold, New York 1985.

[NASA96] Method for Encoding Low Entropy Data, *NASA Tech Briefs*, April 1996, p. 20.

[PEAR85] Pearson, Don E. and John A. Robinson, "Visual Communication at Very Low Data Rates," *Proceedings of the IEEE*, April 1985, pp.795-812.

[RUIU96] Ruiu, Dragos, "An Overview of MPEG2," Hewlett Packard 1996 Digital Video Test Symposium.

[TURK91] Turk, Matthew and Alex Pentland, "Eigenfaces for Recognition," *Journal of Cognitive Neuroscience*, Volume 3, Number 1, 1991, pp. 71-86.

[WOLF96] Wolff, Gerry , World Wide Web at http://www.sees.bangor.ac.uk, 1996.

CHAPTER 2. LOW POWER ELECTRONICS TECHNOLOGY TRENDS

SUMMARY

A paradigm shift is evident in recent developments in semiconductor electronics. More than a specific initiative or an isolated need, there is a general industry-wide confluence of trends that raises the importance of low power electronics and emphasizes the value of a system-level approach to meeting this challenge. Although the objectives vary by product class, low power is a theme that encompasses traditional portable electronic devices and desktop systems, as well as new classes of products now in conceptual design.

Designers of integrated circuits and systems are aware of the impact of power dissipation levels on speed, complexity, size, and cost – the traditional design goals. Higher clock speed and integration levels increase power consumption and waste heat, and increase on-chip electric fields which in turn decrease reliability. Increased power and heat rejection may also lead to higher packaging costs [SING95].

The industry has responded with several approaches, ranging from aggressive advances in state-of-the-art materials and manufacturing processes, to treating low power design as a new (and complex) area for development of highly capable simulation tools, to case by case *ad hoc* solutions that exploit unique requirements or features of the application at hand to reduce power requirements. There is a consensus that, despite the engineering challenges, the investment costs and the hierarchy of various limitations looming on the horizon, that the near term will continue to see substantial progress in development of low power electronics.

1. INTRODUCTION

The remarkable advances in electronics over the past thirty years have come as designers responded to market demand for increased performance, smaller size, lower cost, and greater reliability. For most applications and product categories, minimizing power has not been a key design objective; cooling capacity limits might impose a practical maximum power, but without introducing low power as an explicit design objective. Power level reductions for many product categories have occurred as new technologies were adopted, but the design objectives listed above generally remained unchanged.

In contrast, certain niche applications have always depended on low power as a key asset for success. Personal electronic devices, such as watches and calculators, must be small and light-weight, making low power operation a must. Some military applications and most space flight hardware must operate for extended periods with very limited energy conversion and storage resources.

The emergence of low power electronics (LPE) as an important industry theme today traces back to the convergence of these two on-going trends with a demand for new uses for low power devices that is only now evolving. Current interest in LPE, then, is the result of:

- Extension of the early base of applications in personal electronics and specialized hardware.

- Requirement for continued power reductions to support expanded use of densely packed smaller MOSFET devices.
- A new set of applications, portable and stationary, for which LPE is an enabling technology [TERM95].

The importance of LPE as an industry theme is highlighted by programs of funding agencies that are designed specifically to pursue developments in LPE research, and focused technical meetings sponsored by the Institute of Electrical and Electronics Engineers (IEEE) and the Association for Computing Machinery (ACM). This new emphasis is more than a passing trend in electronics design; modern initiatives in LPE represent a true paradigm shift in electronics design practice. Instead of treating power as an incidental consideration, low power electronics is now a critical system-level design objective: "...what will be needed is a low power culture, where designing for low power is as important and innate as designing for high performance and/or minimum area has been in the past" [TERM95].

The LPE "culture" identified by Terman and Yan has as its objective a system-level design characteristic; therefore, a successful implementation of LPE may include many disciplines related to system architecture, circuit design, design tools, materials and processing, algorithms, and device operation. For every level of technology, an optimal, lowest-power design will probably require simultaneous advances in many of these fields.

How low is low? While it is difficult to give a quantitative characterization of a trend that encompasses so many different applications and product categories, the Defense Advanced Research Projects Agency (DARPA) has as its goal development of "a mainstream technology base for a class of electronic systems that will dissipate less than one percent of the power of current technologies" [LEMN96a] – a reduction of two orders of magnitude over current state-of-the-art. This objective will probably not be satisfied based on advances in a single area; instead, it will be achieved through stepwise improvements in several of the disciplines mentioned above.

That such reductions are feasible has been demonstrated by applying a multi-level optimization approach to the chip set design for a portable multimedia terminal [CHAN95a, CHAN95b]. The result is a design that requires less than 5 milliwatts at 1.1V, which is better than three orders of magnitude below power requirements for similar commercial products that are currently available. The underlying technology trends that will provide the basis for similar improvements in many other application categories are described in the following section.

2. LONG-RANGE TECHNOLOGY TRENDS

In one sense, the current interest in LPE can be seen as the result of forecasting a continuation of long-standing trends in the evolution of microelectronics products, driven by clearly articulated demands of end-users in consumer and industrial markets. But the LPE focus also anticipates entirely new generations of products that are expected to create new demands that will require a low power approach as an enabling capability. This section describes the dominant market trends behind the urgent need for LPE, and catalogs some of the key products or product classes and the expected benefits of LPE.

The industry has responded with a systems-level view of how power dissipation can be reduced in several contributing areas. The section concludes with a discussion of theoretical and practical limits that govern future improvements in highly integrated LPE.

2.1 ANALYSIS OF TRENDS

Three trends in end-user products have developed over the past 5-10 years that bear directly on the importance of LPE in the future. Each trend is a demand-driven need for LPE; together, the breadth of applications they span has forced the need for a true paradigm shift in electronics.

The first is a trend towards increasing power to deliver greater performance in stationary or "desktop" products. The second is a trend towards decreasing power levels to achieve greater convenience and true portability of personal electronics of various types. The third trend, applicable to both stationary and portable products, is for increased packing density in order to increase speed. Note that these directions are driven largely by user demand.

Until recently, designers of traditional digital microelectronics applications, such as stationary desktop personal computers, have not addressed power dissipation as a first-order design goal. Power conservation was important only to satisfy some other objective, such as reducing the cost of the product's power supply and cooling needs, or as a remedial action to solve a power problem that occurred as a consequence of other design choices. As a result, power dissipation increased in many microprocessor, memory chip, and application-specific integrated circuit (ASIC) product lines [MEIN95]. An early crisis in heat rejection was avoided in the late 1970's and early 1980's, when the industry shifted from bipolar technologies to lower-power CMOS as the standard technology. Further reductions in power levels have come since then, as the result of scaling down to smaller MOSFET devices [TERM95].

However, the overall trend has been one of increasing – not decreasing – power draw at the end-user application level. The Semiconductor Industry Association (SIA), in its National Technology Roadmap for Semiconductors, forecasts that this trend will continue, as designers respond to market demand for greater speed and performance. By 2010, required power levels for devices with or without heat sinks will better than double over present levels [NTRS94]. This forecast is based on an extrapolation of Moore's Law [NASA95] and the prevailing trend of steadily increasing clock frequency. However, the increased cost to manage higher power levels conflicts with strong market pressure to constantly improve the cost-to-performance ratio of each product class.

Concurrently, many classes of personal portable electronic devices have required progressively smaller size and weight to satisfy end users; this is the second of the three trends. Beginning with transistor radios, wrist watches and pocket calculators, then suitcase-sized or "luggable" computers, these products now include true portable computers (e.g., laptops) and cellular telephones. Each new generation or class must meet user requirements for convenience and portability. To succeed, a design must combine low power requirement with a small, light weight battery to produce a reasonable operating time before recharging is needed. Optimal designs have pushed both LPE and battery technology to improve product performance and life, while reducing cost, weight, and size.

Finally, many products, both desktop and portable, require continued increases in speed and performance, which implies higher packing densities that make power dissipation a key issue. Very high performance demands are placed on back-end processors such as database machines and servers that must support many users simultaneously. The result is that LPE becomes a primary design requirement for these high-performance applications.

The following table summarizes these three market-driven trends that all point to the importance of low power now and in the immediate future [MEIN95]. Together, the product classes affected by the three trends constitute a majority of current end-user market applications, thus emphasizing the importance of the low power goal that all three have in common. Although portable products were the first class to identify low power as a key requirement, they remain the most difficult product class to satisfy. It is clear that LPE is an important fundamental design principle that goes far beyond support of portability.

Table 2.1 Demand Sources for Low Power Electronics

Market Need	Scope	Design Implications
Competitive cost-to-performance ratio for desktop equipment	Broadest need, affecting greatest number of end-user products and product classes	Reduce cost of power supply and cooling by reducing power draw
Small, long-life portable devices	Earliest stated need for LPE, most demanding design	Meet demand with novel LPE designs and small, long life batteries
High-end system performance gains	Most recently articulated need includes both desktop and portable product lines	Reduce power dissipation to compensate for increased packing density

As the industry has recognized the broad need for LPE identified by the product trends above, two significant changes in research directions have occurred. The first is that dedicated conferences, publications, and initiatives have increased the visibility of LPE issues throughout the electronics industry. This movement has gathered momentum only in the last five to six years. There are now scheduled conferences that specialize in low power electronics. For example, the ACM Special Interest Group for Design Automation, the IEEE Solid-State Circuits Council and the IEEE Circuits and Systems Society have joined to cosponsor the International Symposium on Low Power Electronics and Design. In 1995, this symposium consolidated two other events that had previously been organized and held separately. Since 1991, the European project, Power And Timing Modeling Optimization and Simulation (PATMOS), has held an annual workshop at a university in Germany, France, or Spain. PATMOS covers design and computer-aided design (CAD) issues, and dedicates several sessions to various aspects of low power as it relates to electronics design and design tools.

Equally significant in the research community, DARPA has initiated a Low Power Electronics Program to focus LPE research which is of interest to several other DARPA offices and programs [LEMN96a]. One stated objective for this program is to foster development of a new electronics technology base for hand-sized information systems [LEMN94]. Several universities sponsor active research programs in LPE and closely related fields. These include UCLA, UC - Berkeley, USC, MIT, and the University of

Illinois, among many others. The PATMOS community is comprised largely of researchers from European universities.

As one example of the research programs now in place, the adiabatic charging MOS (ACMOS) group at the Information Sciences Institute, University of Southern California is exploring adiabatic charging and pulsed-power techniques for building low and ultra-low power circuits that recycle energy instead of shedding excess as heat. These concepts are claimed to have interest for CMOS chips, liquid crystal display (LCD) panels, and micro electro-mechanical systems (MEMS) electrostatic actuators [ACMO96].

Since digital CMOS circuit design is the focus of the LPE effort, discussion in this chapter centers on digital devices and circuits. However, a second example is worth mentioning, because it focuses on a class of devices – operational amplifiers (op-amps) – frequently included in analog systems. Op-amps are included in many analog systems as part of larger integrated circuits, e.g., in switched-capacitor filters or data conversion circuits. MIT is conducting a study of op-amps under the constrained power levels that would be encountered in LPE designs [CHOI96].

The second change in research direction is a general recognition among semiconductor industry researchers and designers that a systems approach is required to successfully address the challenges of LPE. This implies that the design begins with low power as an explicit objective, and that a multiplicity of techniques may be called upon to produce a low power design. Four basic strategies have been identified [PEDR96], which will be explored in greater detail in subsequent sections of this chapter.

- Reducing chip and package capacitance. Materials and manufacturing process developments such as CMOS scaling to submicron device sizes, silicon-on-insulator technology, and advanced interconnect substrates are some of the promising approaches. Though effective, these methods require up-front investment, and must deal with a number of process-related issues before full potential can be delivered to production.
- Scaling the supply voltage. The recent trend to progressively lower supply voltages continues. This approach requires new fabrication processes for integrated circuits, and may require other system-level changes to incorporate low voltage chips.
- Better design techniques. A relatively low initial investment in advanced CAD tools may produce significant power reductions, although lead time is needed to develop and test the tools. The tools also require improvements in power estimation capability.
- Power management strategies. The designer can take advantage of application-specific characteristics to design a low-power solution.

The coordinated systems-level approach has also brought focus to the importance of limiting factors that control future LPE developments. J. D. Meindl has organized these into a hierarchy of limits [MEIN95, NASA95]:

- Fundamental limits derived from the laws of physics. These apply to all materials, process technologies, and designs:
  - A thermodynamic limit, the minimum usable signal power due to thermal noise

  - A quantum-mechanical limit, the minimum signal power according to Heisenberg's Uncertainty Principle
  - The speed of light limitation on signal propagation.

- Material limits associated with properties of specific materials, such as thermal conductivity, carrier mobility, carrier saturation velocity and, for any material used as an interconnect, the relative dielectric constant of its insulator. Material limits apply to all devices, regardless of size.
- Device limits on transistors and interconnects, such as the minimum effective channel length of a MOSFET. Device limits are independent of the circuit design.
- Circuit limits on design for low power:
  - Supply voltage and threshold voltage must be selected to ensure signal quantization
  - Energy dissipated per switching transition is a function of physical capacitance and supply voltage
  - Intrinsic gate delay impacts power dissipation
  - Response time of global interconnect circuits is determined by circuit limits.
- System limits associated with overall chip architecture, the power-delay product (energy) of the implementing CMOS technology, the heat rejection capacity of the chip, and the physical size of the chip.
- Practical limits imposed by manufacturing techniques, and the expected return on capital investment in the manufacturing plant, among others.

2.2 FUTURE APPLICATIONS

This section presents a brief survey of some of the products and product classes most affected by the need for low power electronics, and describes likely near-term future evolution of these products. Three generic design points have emerged to push development of commercial low power electronics: portable devices that operate on very small power budgets, portable computers designed for ever higher performance, and stationary systems that do not operate on battery power.

Micro-power battery operated portable devices include hearing aids, implanted medical devices such as heart pacemakers, pagers, cellular telephones, and the much discussed personal digital assistant (PDA). Power levels and design requirements vary considerably over this range of products. Very low operating power and a battery lifetime measured in several years is typical of implanted medical devices. Small size is also a critical design requirement for invasive devices, as well as hearing aids. Pagers, now pocket-sized, are evolving to two-way messaging capability, which will require reductions in component sizes to maintain the same general product size. Portable cellular telephones and their batteries continue to decrease in size and weight, even as time between recharges increases. Even hand-held electronic games have benefited from LPE progress: sophisticated games are indeed enabled by low power. The next generation of game machines is predicted to use a 10 MIPS microprocessor at a cost of $10 and less than 1 W power consumed – a processor speed that took $10 M, 10 kilowatts, and lots of space to deliver in the mainframe of 20 years ago [STOR95].

The PDA is now more concept than real product. More than a portable personal computer, more than a wireless communications device, the PDA concept is a multi-media terminal with wireless access to a global communications network for digital transmission of voice and data. The PDA handles location-independent communications, image acquisition and display, on-demand access to databases, and interactive computing.

Current products are called PDA's by their vendors, but fall short of this idealized concept. As one example, U.S. Robotics offers Pilot (about $370) – a 5.5 ounce hand-held computer that runs on two AAA batteries. The 3.4-inch monochrome screen displays icons to control functions such as date book, note-taking, address and phone book, and to-do lists. Data can be entered by using a stylus to type on a keyboard displayed on screen, or by using the handwriting recognition software included. A cradle attaches to a desktop personal computer serial port for bi-directional data transfer with Pilot [PERS97].

Pilot is certainly an impressive state-of-the-art product for its size and price, but it is not the complete PDA concept described above. To achieve the full potential of the PDA concept requires further advances in highly integrated electronics, display technology, multi-media advances for both input and output, and a low power approach throughout. Human interface capability must expand to include speech recognition and synthesis. The speech recognition function now requires a full board and as much as 20 W of peak power [STOR95]; the PDA implementation will have to fit on one or two chips. (There is also the not-insignificant matter of the global standardization needed to support interoperability of PDAs, but that is beyond the scope of this discussion.)

Portable computers have steadily increased market share of the sales of all personal computers. In the last 10-15 years, suitcase-size and briefcase-size portables have been replaced by notebook and sub-notebook products. Each generation is smaller and lighter, includes more capability and higher performance chips, and provides longer battery life. Battery life, in particular, is a current focus of intense development effort, concerned with both the power supply's characteristics and the power consumed by various subsystems.

As a design point for discussion, a current (as of 1993-4) typical notebook computer with a 25-33 MHz microprocessor, an internal hard disk, and a monochrome liquid crystal display would be equipped with a 30 watt-hour battery, and would dissipate about 8 W in operation, giving a battery life of three to four hours of continuous operation before recharge. Of the total power required at full operation, the display takes about 2 W, the hard disk about 1 W, the CPU 1 W, the video chip-set 1W, memory and other CMOS logic an additional 2 W, and 1 W is dissipated by DC-DC losses [HARR95]. For a color display, the power requirement jumps to about 20 W, of which over half is budgeted for the display. Harris forecasts that products to be available in the next year or two will reduce this to 5-6 W total for a color notebook system. This forecast assumes "an aggressive low-power design strategy" in the product's development.

Future products will also increasingly offer some form of wireless communications capability. While the display's power requirements will continue to dominate the total, power for communications will become an increasing fraction of the product's budget. Whether the use is cable replacement, local area, or wide area communications, low power radio frequency and/or low power infrared technologies are expected to evolve to meet power objectives for advanced portable computers.

Though most of the industry focus on LPE is driven by the types of portable products just described, there is also a continuing need to reduce power consumed by stationary systems. These include desktop computers, high-end workstations, servers, database engines, and modern generations of mainframe computers, as well as telephone switching systems installed on customer premises. The power limits may be imposed by cooling requirements for chips or other system components, by site requirements, or even by so-called "green laws" regulating the office environment. Much of this equipment is now installed in standard office space, without the environmental controls formerly available in the build-out of a dedicated equipment room or computer center.

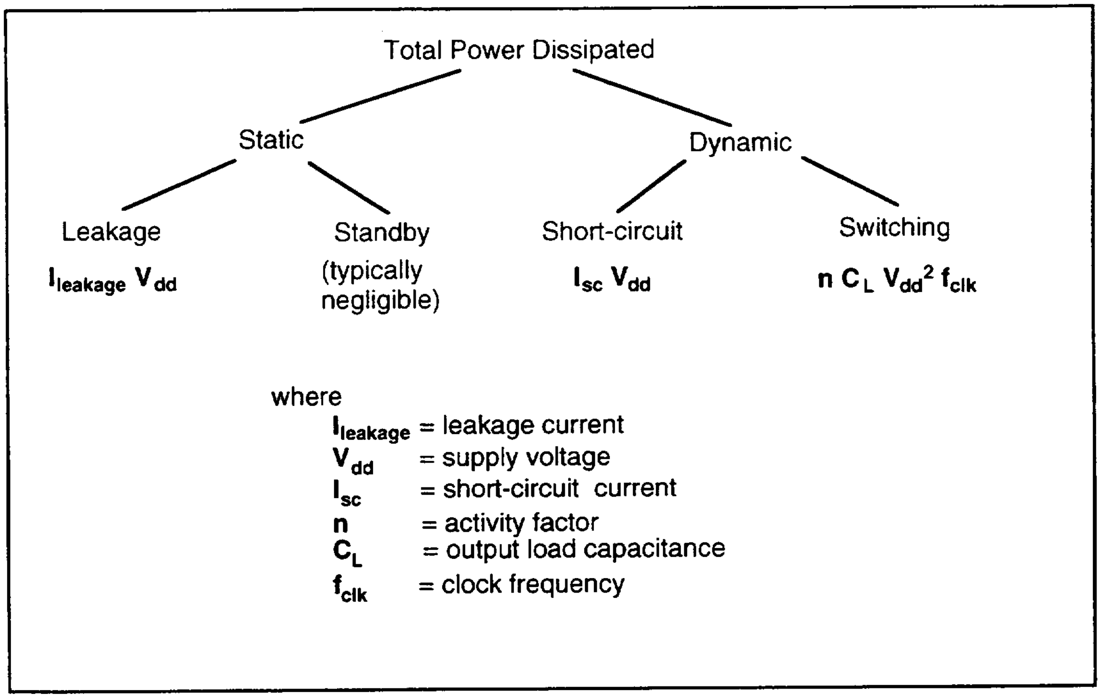
Beyond commercial applications, there are also future space applications of LPE that are of vital importance in future civil and military space programs. The unrelenting pressures of decreasing real budgets has forced program planners to adopt a new paradigm for spacecraft and mission design. A prime example of the important role to be played by LPE is the Discovery series of planetary missions. Originally proposed as a means of filling in the gaps in planetary exploration and encouraging new research between launches of major Galileo-class spacecraft, Discovery missions now represent all of the most recent new start approvals given by NASA. The "better, faster, cheaper" approach almost inevitably also means smaller as well.

Small spacecraft have been studied in various planning contexts for some years (for example, [JONE89]). Now, extremely small size is one of the key goals of the New Millennium program of spacecraft system and subsystem technology development at the Jet Propulsion Laboratory (JPL). Science instruments and imaging devices are also the focus of continued attempts to reduce size, weight and power consumption. An example of such work is the "camera-on-a-chip" reported by JPL [FOSS95]. Implemented in standard bulk CMOS technology, the active pixel sensor design integrates the detector array with on-chip timing, control, and signal-chain electronics. Analog-to-digital conversion is also included. Developments such as this could be employed on future generations of surface rovers even smaller than the Sojourner rover on the Mars Pathfinder spacecraft.

LPE design goals may also be helpful in improving reliability of future generations of in-space electronics. As highly scaled devices have been deployed in operating satellites, hard errors have occurred as the result of cosmic ray hits. This damage has been observed in DRAM's scaled to 0.8 micron feature size. However, a theoretical scaling result has been reported indicating that, for feature sizes below 0.5 microns, a device optimized for low power will experience much lower hard error rates than one designed for high speed [JOHN95].

2.3 EMERGING TECHNOLOGIES

To understand the many technologies and approaches now being pursued to reduce power, it is instructive to consider the sources of power draw in typical CMOS circuits. The following figure organizes four sources into static and dynamic, and indicates for each the design variables that influence power level. In turn, the designer can control these variables in several ways to reduce the total power dissipated by the circuit.



where
$I_{leakage}$ = leakage current
V_{dd} = supply voltage
I_{sc} = short-circuit current
n = activity factor
C_L = output load capacitance
f_{clk} = clock frequency

Figure 2.1 CMOS Power Dissipation Sources

Static power dissipation is the sum of the leakage and standby dissipation. Dynamic power dissipation is the sum of the short-circuit and the switching (or capacitive) power dissipation. Of the four, switching power is the most significant source to be managed by LPE techniques. The following paragraphs discuss each of the four components, and indicate the motivation for various power-minimizing techniques which comprise the overall systems-level approach. See [PEDR96] for a more complete discussion of dissipation sources.

Leakage power is the product of the leakage current and the supply voltage (V_{dd}) so minimizing leakage power requires control of at least one of these parameters. In turn, the leakage current ($I_{leakage}$) is determined by the fabrication technology employed, and by the transistor threshold voltage (V_t). Typically, $I_{leakage}$ can be reduced to negligible values by making the right design choices.

Standby power is shown in the figure as a negligible quantity. It is the product of supply voltage and any standby DC current which is drawn continuously from the supply voltage to ground. Although standby currents are significant in a few special-purpose designs such as memory cores, they are small enough to be ignored for most CMOS technologies, logic styles, and circuit structures [PEDR96].

Recent work has shown that the third component, short-circuit power dissipation, which is the product of the short-circuit current (I_{sc}) and the supply voltage, depends on device design and performance parameters, such as transistor size, load, and input/output ramp times. With judicious design selections to control I_{sc}, short-circuit power can be held to 10-15 percent of the total dynamic power in most cases [CHAN96]. The exception would be a design for very high performance, where a short-circuit power penalty is inherent in the design choices taken to maximize performance [PEDR96].

The fourth and most significant of the components of power dissipation is the switching or capacitive power dissipation. Switching current flows to charge and discharge capacitive loads during logic changes. From the equation in the figure above, switching power depends upon the activity factor (n), the output load capacitance (C_L), the square of the supply voltage, and the clock frequency (f_{clk}). The activity factor is the average number of output transitions per unit of time, and is determined by the power management scheme selected by the designer [LEMN96b].

2.3.1 The Trade Space For Low Power Design

From the preceding discussion it is clear that low power design must focus on the dynamic components of power dissipation, since the static components are easily treated in most cases. Since short-circuit power consumption is usually less than 10-15 percent of the total dynamic power, that leaves switching power as requiring the most attention for LPE design. Inspection of the equation for switching power dissipation shows that there are three variables that determine power dissipation: voltage level, physical capacitance, and data activity (the last being a combination of the activity factor and the clock frequency). This section is a short summary of the trade space for low power design, condensed from [PEDR96].

Voltage Scaling

Since the switching power varies as the square of supply voltage V_{dd}, reducing the supply voltage is an effective way of reducing power. For example, cutting supply voltage in half reduces power dissipation to one-fourth of the original level, if other quantities are constant. Reducing V_{dd} also reduces power dissipation throughout the entire design. The drawback to this approach is a speed penalty that grows in severity as V_{dd} approaches threshold voltage, V_t. Threshold voltage can also be reduced, but there are limits as to how far. For operation of CMOS circuits at room temperature, the practical limit is reported as 0.3 V [DAVA95]. Since delays increase as V_{dd} approaches within two to three times V_t, a practical target is to reduce supply voltage to just under 1 V.

Physical Capacitance Reduction

The power dissipation in a circuit depends on the capacitances seen by individual gates and on the interconnect factors in the circuit. Therefore, power in a circuit can be minimized by using less logic, smaller devices, and fewer and shorter wires [PEDR96]. There are limits to each of these approaches to reducing capacitance. For example, using smaller devices lowers the physical capacitance, but also reduces the speed of operation of the circuit.

Switching Activity

Switching activity in a circuit draws power. Minimizing switching activity is accomplished by controlling the clock rate, which gives the average periodicity of data arrivals to the circuit, and by controlling the activity factor, which gives an expected value of the number of state transitions generated in each clock period. The obvious trend of ever-increasing clock rates implies a corresponding increase in the switching activity per unit time, and an attendant increase in power. Glitching, the spurious unwanted transition activity that may occur in a circuit design, increases the switching activity and, therefore, the power consumed.

As the discussion above indicates, no one of the three factors can be minimized independently, since there are other consequences, besides the intended one, that affect performance. The general problem is one of finding an optimal design, taking all factors into account. But it is not possible to specify one objective function for this optimization; the choice often depends on the application. Pedram [PEDR96] cites some examples:

- Minimize energy if extended battery life is the key, and speed is less important.
- Minimize action (the product of energy and delay time) if battery life and speed are important. The problem here becomes finding the largest reduction in energy for the smallest performance penalty.
- Minimize energy subject to a delay constraint if timing is specified at the system level to respond to user requirements.

Typical objectives for a design are actually somewhat more complex for many products, since they respond directly to market forces. For example, extended battery life may be coupled with requirements to keep weight and packaging costs as low as possible.

The sections that follow describe some of the specific approaches now being used to achieve design and operation of low power circuits. Each attempts to control one or more of the three variables described above.

2.3.2 Scaling Of CMOS Devices

Device scaling has progressed steadily, following Moore's Law, as illustrated Figure 2.2, which is reproduced from Figure 26 of [MEIN95]. Average minimum feature size is plotted versus time, including an extrapolation to 2030 (again relying on Moore's Law). This trend has continued for the past 35 years because the cost per logic circuit has continued to decline, continuing to support the march towards ever-smaller feature sizes.

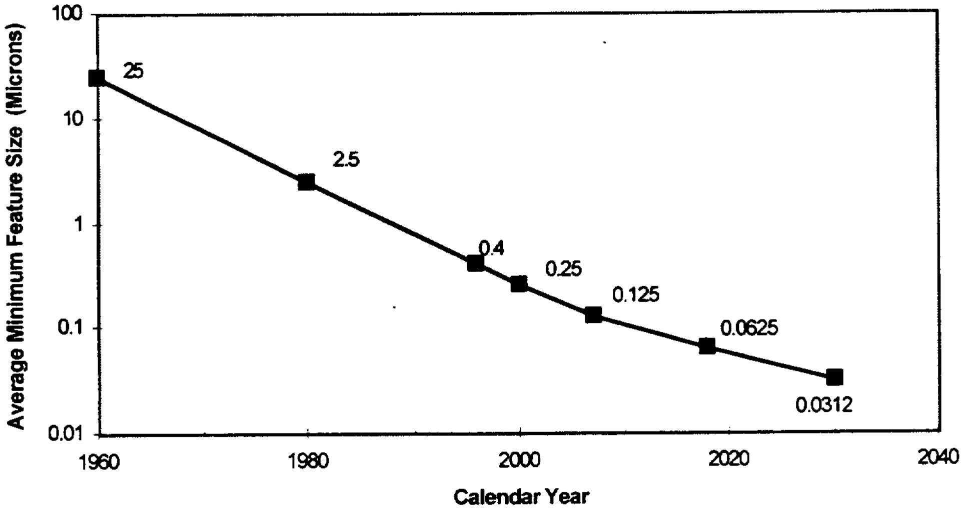


Figure 2.2 Evolution of CMOS Feature Size

In its simplest form, scaling a device means that every linear dimension is reduced by some constant scale factor, k. This reduces the size of the MOS transistors and wires by the same factor as well as reducing the applied voltage, preserving characteristics of the electric-field, but in a smaller device. For the new device, density increases by a factor of k^2, speed increases by a factor of k, and power dissipation is reduced by a factor of k^2 because both voltage and current are reduced. Since power density remains constant in this scenario, more logic circuits can fit into a given area with no increased power dissipation [DAVA95].

However, constant electric-field scaling such as this requires departing from standardized supply voltage levels, which has generally been avoided in past design practice. One team implemented a 5 V to 3 V interface I/O cell in an ASIC chip level design for an advanced tape drive controller to take advantage of power reductions at 3 V without changing the rest of the 5 V system [SCHM96]. Davari et al. describe a more generalized approach, introducing a multiplier that allows the device designer to control scaling of the supply voltage somewhat independently of the change in linear dimensions. But they conclude that part of the new paradigm of LPE design is the inevitable move to more frequent voltage reductions – to 1 V or less.

This trend is contingent on the assumption that the cost per logic circuit continues to decline throughout the forecast period, and that technology improvements will continue to support the trend. For instance, the wavelength of deep-ultraviolet light now used in the photolithographic process is about 0.1 micron [NASA95]; to stay on the curve shown in the figure, new generations of photolithography or X-ray lithography manufacturing must be deployed between 2005-2010. In turn, this requires a vigorous expansion in the size of the market to warrant the investment in new technology.

2.3.3 SOI Devices

Silicon-on-insulator (SOI) is a semiconductor material manufacturing technology in which metal oxide semiconductor devices are built in a thin silicon film grown on an electrically insulating substrate, as illustrated in Figure 2.3. SOI is a silicon wafer to which an insulating layer is added between the thin silicon film and the thicker silicon substrate [LEMN96b].

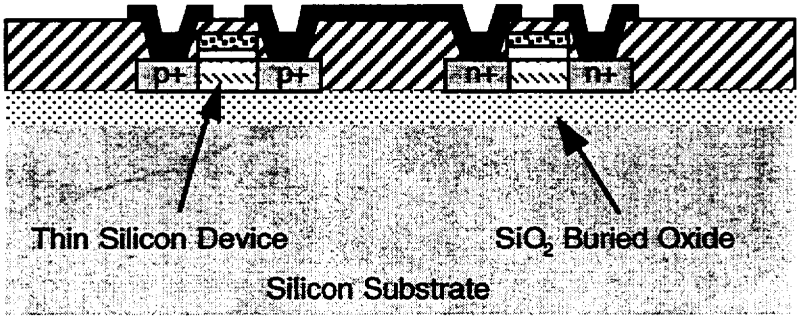


Figure 2.3 Cross Section of an SOI Transistor

As compared to conventional bulk CMOS, SOI technology reduces parasitic, device, and interconnect capacitances. Some SOI implementations may also reduce the threshold voltage. In turn, these advantages of SOI may be used by the designer to impact power dissipation and device performance. Davari, et al. report that circuit designs implemented in CMOS using SOI have been shown to produce performance gains of 1.5-2.5 over CMOS on bulk devices that use the same lithography. [DAVA95] Since isolation properties are also improved, circuit density can be increased, leading to the possibility of a simultaneous increase in speed and power [STOR95].

Figure 2.4, reproduced from the Stork paper [STOR95], illustrates the performance gains that may be realized with SOI. The figure, which summarizes modeled results and not laboratory observations, plots clock frequency versus power for three different feature sizes. Feature size is represented by the gate length, and includes the current 0.7 micron technology, as well as two extrapolations to smaller sizes. For each size, a pair of curves is plotted to show the performance gain achieved in moving from Si to SOI. The five points on each curve were generated using the same assumptions, including supply voltages as marked on the figure.

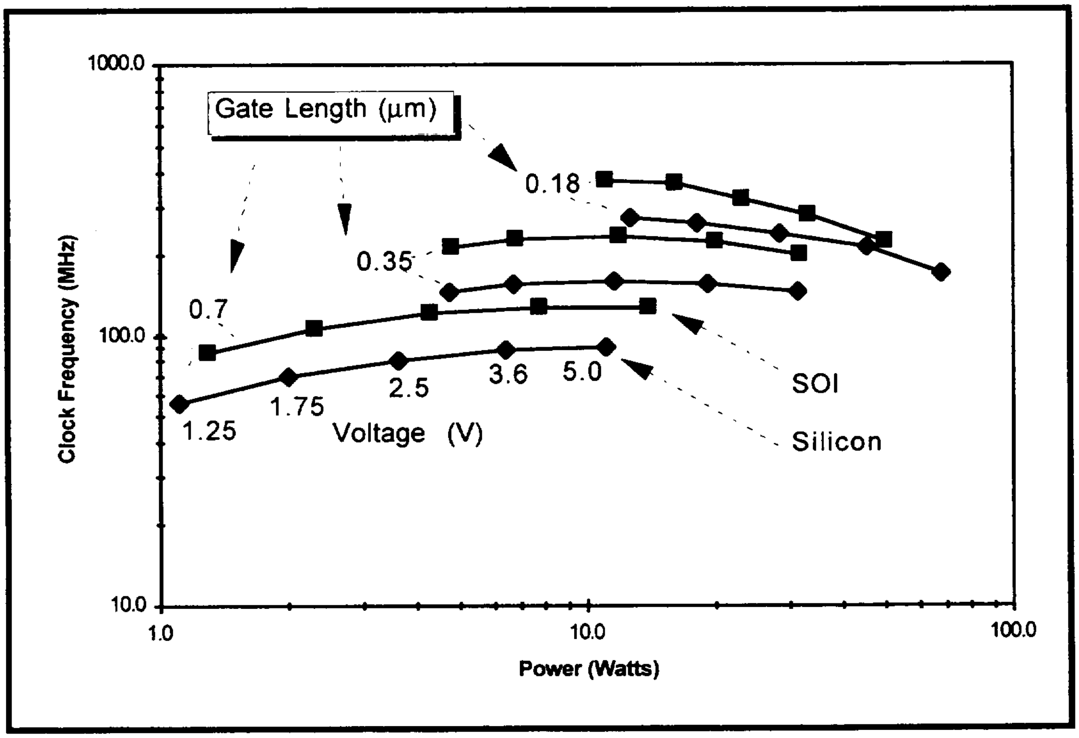


Figure 2.4 Comparison of SOI and Silicon Technologies

For current technology (0.7 µm) and any specified power level in the domain, the model estimates that switching from conventional bulk CMOS to SOI increases the clock frequency approximately 1.5 times, which falls within the range quoted above. For example, comparing the two curves at a power of 5 W, the performance gain is from 80 MHz to 120 MHz. This pattern is repeated in future generations (0.35 and 0.18 µm). Over all three generations portrayed, and for either Si or SOI technology, the clock frequency is maximized at a power level slightly above 10 W; however, the optimum shifts from the highest power level at large size to the lowest power level at smallest size. At today's technology, power can be reduced while clock speed stays nearly constant. This potential leverage seems more difficult to achieve in later generations, since both supply and threshold voltages have already been reduced.

SOI wafer substrate can be formed by any of several manufacturing processes, of which there are two leading candidates at present. The first is separation by implantation of oxygen (SIMOX). Oxygen atoms are implanted in a silicon epi-wafer which is then heated to form SiO_2 and anneal the wafer. The second process, bond and etch-back of SOI (BESOI), bonds two wafers at high temperature, then etches back one of the wafers. Properties of the SOI wafers produced may depend upon which process was used to manufacture them. Furthermore, neither of the processes is cheap enough or mature enough to consistently produce large wafers of acceptable quality in the volume required by commercial applications.

While there are diverging views on the cost and time needed to address these issues, there is agreement that there are no fundamental show-stoppers in the way of achieving the promise of SOI technology. SOI builds on the substantial silicon technology and infrastructure investment. Moreover, SOI has been in use for some years in specialized applications such as building integrated circuits for high voltage, high temperature, or radiation-hardened systems. SOI also improves noise coupling in mixed digital/analog signal uses.

2.3.4 Low Power Memory Development

Random access memory (RAM) is one of the major areas in semiconductor electronics that has seen substantial progress in reducing power. This trend has been evident for some time, as active power has been reduced with every new generation of memory chip. Furthermore, these advances have come even as supply voltage remained constant, chip size increased, and memory access speed improved. Progress has occurred for both dynamic RAM (DRAM) and static RAM (SRAM).

During the last decade, DRAM power dissipation has been reduced by two to three orders of magnitude for a chip of fixed memory capacity. For example, a hypothetical 64 Mb DRAM design implemented in 1990 CMOS technology reduces power draw by two orders of magnitude over the NMOS technology of 1980, and state-of-the-art technology would reduce power to about one-tenth of the 1990 level. [ITOH95] Lower charging capacitances and reduced operating voltage are the keys to these advances. The prospect of reducing supply voltage to 1.5 V or lower suggests further gains for DRAM chips of 64 Mb and larger; already, a 64 Mb DRAM at 1.5 V has been reported operating in the laboratory. Reducing the power level required for data retention could be significant for battery backup use, since less expensive DRAM could be substituted for the SRAM now used for that role. This could also have implications for the capability and cost of hand-held communications equipment.

SRAM development focused on low power operation beginning with the earliest generations of product. Design emphasis was placed on achieving low power levels for standby and data-retention, concurrently with increasing memory size. Recent work in SRAM development has emphasized high speed for applications, such as microprocessor cache memories. The low power design problem differs from DRAM's, because SRAM doesn't require a refresh operation to preserve memory contents, and because SRAM static current dominates total active current. Therefore, although static current, voltage and capacitance are the keys to reducing power dissipation in both cases, the emphasis for DRAMs is on capacitance, while the emphasis for SRAMs is on static current.

2.3.5 Low Power Radio-Frequency Circuit Design

Low power integrated circuits designed for use in the radio frequency (RF) front-ends of portable wireless products have made substantial reductions in product size and power consumption. These products include FM radios, pagers, and cellular telephones. The pager network now covers many parts of the world, and two-way paging is under active development. The latest cellular phones, scarcely larger than a pager, fit easily into a shirt pocket. The transition from traditional analog communications transceivers to digital signal processing methods has sparked the renewed interest in advances in radio telephony. A survey article by Abidi [ABID95] predicts that future digital cellular telephones will include single chip silicon transceivers and architectures that reduce passive components. All-CMOS designs for radio transceivers in the 900 MHz to 2 GHz bands are promising. Some of these wireless products, such as pagers, can implement well designed protocols with explicit power reduction features [MANG95].

2.3.6 CAD Tools And Methods For Low Power Design

As discussed in Section 2.1, one of the four strategies for low power now being pursued by the electronics industry is better design techniques. While there is no consensus as to which of the four strategies will produce the most progress, it is generally agreed that the industry must continue to pursue all four in the near term. The promise of pursuing LPE via improved design methods and tools is twofold: the perspective and impact is at the system or product level, and the investment is small as compared to the changes in materials and manufacturing processes described in earlier sections. However, the road to improved methods and tools is not without pitfalls. CAD tools for electronics design at various levels of abstraction are difficult to develop and use because of the nature of the problem they address.

The goal of CAD methods is to optimize one or more system values that are relevant to the application at hand, for example, to minimize power dissipation, subject to various constraints. But the first problem is to make a useful *a priori* estimate of the expected power dissipation. Recall from Figure 2.1 that the dominant contribution to power dissipation in CMOS circuits is the switching or capacitive power, which is the product of several terms described previously:

$$\mathbf{P = n\, C_L\, V_{dd}^2\, f_{clk}}$$

Two of these factors are problematic for estimating power dissipation [SING95]. Detailed knowledge of the implementation is required to calculate an accurate value for physical capacitance, $\mathbf{C_L}$,. This is the easiest of the two, since the problem is at least deterministic. In contrast, the switching activity factor, **n**, is in turn dependent on several factors, some of which must be expressed as stochastic variables. This introduces uncertainty and added complexity to the process of power estimation. Different classes of solution methods have been proposed for power estimation, including simulation methods, stochastic models, library-based models, and models based on information theory [PEDR96].

Power minimization approaches can be implemented at any of several different levels of design abstraction, from the highest levels (behavioral and architectural) to the lowest levels (logic and physical). In general, the higher the level, the greater the potential impact any power-saving steps will have on the overall design. At the behavioral and architectural levels, the problem is often a trade-off among many contradictory design parameters whose impact can only be approximated at the earliest stages of the design process. System-level specifications may therefore be fixed based on preliminary information, limiting the possibilities for power minimization [SING95]. On the other hand, the lower the level, the better the accuracy of the design. Low-level circuit simulators such as SPICE are commonly used for detailed design work but are time-consuming to run, making trade studies of complex designs too expensive and slow to justify. Moreover, power savings at lower levels tend to be limited.

Wolfe [WOLF96] provides a case study of the practical problems associated with using CAD tools for low power design. The example is an embedded system: interface electronics for a touch-screen add-on product for personal computers. Wolfe, himself a developer of embedded system CAD tools, notes that many designers do not use these CAD tools (as compared to their common use in VLSI work) because they believe that the current tools don't address the critical system-level issues for embedded systems. Indeed, the case study presented was unable to make effective use of system-level CAD tools for exploratory power estimation and optimization. The key reasons given were: the tools do not model the analog components that often dominate low power design decisions; and, the problems that commonly occur at "boundaries" (i.e., between analog and digital components, or between hardware and software implementation) are not modeled.

As a result of the recent interest in LPE, tool designers are beginning to look at big-picture CAD systems. At the register transfer and logic levels of abstraction, the Power Optimization and Synthesis Environment (POSE) has been created to provide a unified power estimating environment to support a global low power design. Another interesting approach is to make available via the Internet a high-level design exploration tool. PowerPlay [LIDS96] manipulates power, timing, and area models of functional blocks and shares design results among users via on-line libraries. Results are presented in a simple spreadsheet format that focuses attention quickly on the biggest power consumers among all system components. PowerPlay is available at the web site noted in the reference.

Some designers remain doubtful about the potential design improvements claimed by the proponents of better design methods and tools as the best investment for LPE. For the skeptics, "low-power design depends on high-power designers." The debate serves to emphasize the difficulty of the underlying problem; circuit design was already complex before low power was added as another dimension.

2.3.7 Batteries For Low Power

Throughout the evolution from vacuum tubes to micro-circuits, advances in battery technology have been linked to advances in electronics. As the scaling down of electronics size and power requirements spurred demand for new products, major improvements in battery performance characteristics have followed. However, battery technology advances at a different (and generally much slower) pace than electronics. In some cases, advanced batteries have reached production readiness only to wait for electronics devices to be designed and built to match the battery performance characteristics. Environmental concerns about the uses and disposal of commonly available products have become an important influence in development and selection of new battery technology.

Battery performance is described by several key measures. Most of these apply to both primary (single discharge cycle) and secondary (rechargeable over several charge/discharge cycles) batteries. The key, of course, is the energy content; but this can be expressed in several ways, each of which may be of particular importance for certain applications. Rated capacity is quoted in ampere-hours for a specified rate of discharge. The cell's power density and energy density are measures per unit volume, often of importance for small portable electronics. Specific power and specific energy measure cell performance per unit mass. Shelf life describes how long the battery can be stored without self-discharging beyond a specified percentage of initial usable capacity. Cycle life is the number of charge/discharge cycles a secondary battery can sustain before falling below a percentage of initial capacity. Other factors such as manufacturing cost and freedom from leakage are also important. Selection of the right battery according to these performance measures depends on the application. The power level may range from a few microwatts for a wristwatch to tens of watts for modern generations of portable computers. The battery's performance depends on its chemical constituents, physical component form (e.g., electrode configuration), operating voltage, size, discharge rate, and duty cycle [POWE95].

The venerable carbon-zinc (Leclanche cell) battery, having improved shelf life and reduced leakage problems, leads consumer sales of batteries world-wide. Alkaline cells, with higher energy density and longer shelf life, dominate the US market for consumer batteries. Alkaline cells also deliver more power than carbon-zinc, which is important for personal devices such as "walkman" tape and disk players, and automatic 35 mm cameras. Several miniature "button cell" batteries are available: zinc-air (oxygen from the air is drawn in to use as the cathode reactant), silver oxide, mercuric oxide, and manganese dioxide cells come in a variety of sizes, shapes, and capacities. The lithium iodine cell, a low power but high energy density primary battery, has been used for more than 20 years to power implanted heart pacemakers. Other lithium cells with liquid cathodes made of sulfur dioxide or thionyl chloride have a higher energy density and rate capability, as well as improved low temperature performance than solid cathode designs; these cells are used in specialized military applications.

Secondary, or rechargeable, batteries are used in those applications that require more power than can be supplied economically by the continual replacement of primary batteries. Applications that have produced a 20 percent growth rate in demand for secondary batteries include cellular telephones, portable computers, and video camcorders. Battery technologies include new designs of sealed lead-acid cells, nickel-cadmium (NiCd) cells with improvements in charge retention and "memory" effect of partial discharge cycles, nickel-metal hydride (NiMH), and lithium ion cells. NiCd cell production is still expanding because of its utility and cost, despite containing toxic cadmium – an environmental concern in use or disposal. NiMH batteries are interchangeable with NiCd in all applications, and provide higher capacity for a given size, although issues remain concerning management of charge/discharge cycles. Lithium ion batteries offer longer cycle life and higher energy density, although requiring active charge/discharge cycle management and somewhat higher cost. Several other promising cell chemistries for secondary batteries continue to be researched.

Environmental concerns surround the development, use, and disposal of both primary and secondary batteries. Many contain toxic metals such as cadmium, lead, and mercury (now outlawed). Some contain flammable material or pose a safety hazard during operation. The risks associated with widespread use and disposal of these products have prompted regulation at many levels. In turn, the industry has redirected research into various battery chemistries based upon environmental impact, and has instituted recycling programs such as the highly effective recycling of lead-acid automotive batteries.

3. RESULTS

3.1 IMPACT ON NASA

The potential of LPE has broad significance for future NASA spacecraft and programs. That potential is timely, in light of increasing pressures to reduce cost and size of flight systems. Any development similar to the "camera-on-a-chip" cited earlier [FOSS95] that reduces on-board space and power requirements directly supports the Agency's current development direction. The Space Science Enterprise's New Millennium program sponsors a new technology initiative focused on MEMS. As "nanosatellite" concepts move closer to reality, all flight subsystems will be reduced in physical dimensions. To deliver data at the increasing rates demanded by investigators requires full utilization of low power memories and electronics [ROBI96a].

"Power is everything!" said one of the NASA engineers in the movie Apollo 13. Planetary missions face another problem that makes low power advances particularly timely. The use of radio-isotope thermoelectric generators (RTG) for on-board power has become prohibitively expensive in the context of lower cost projects, and is doubly so since the Department of Energy no longer shares the cost of development. Moreover, any use of nuclear energy sources is problematic at best. Without RTGs or some remarkable improvements in solar cell efficiencies, missions to outer planets and planetary surfaces are severely constrained. JPL has done recent work on a Power Stick concept that combines several radio-isotope heater units into a flashlight-sized package to deliver fractions of a watt of power. Whether the Power Stick or some form of advanced battery is the power source, an outer planet mission that flies without an RTG requires advances in LPE to support reasonable exploration objectives.

Application-specific integrated microinstruments (ASIM) embedded in ground-based remotely deployed equipment will support condition-based maintenance of the system, provided that very low power wireless telemetry technology is available to support the application. [ROBI96b] This may have use for NASA tracking or other operational systems, and may also be applicable to future exploration of planetary surfaces, such as Mars.

3.2 TECHNOLOGY ROADMAP

The following is a projected timeline for the emergence and development of technologies related to low power electronics. Unless noted otherwise, the forecast numbers below are taken from the National Technology Roadmap for Semiconductors [NTRS94]. Of the four basic strategies for LPE identified in Section 2.1, the biggest near-term gains are likely to come from reducing capacitance (through improvements in material and processes) and scaling voltage. Better CAD design techniques seem likely to require more time for development to produce dramatic reductions in power dissipation.

1996-1998: DRAM reaches 256 Mb per chip; supply voltage for desktop systems reduced to 2.5 V, with battery-supplied systems at 1.8-2.5 V. Minimum feature size for DRAM is 0.25 microns.

1998-2000: DRAM at 1 Gb per chip; performance-optimized clock speed is 600 MHz, supply voltage is 1.8 V for desktop, 0.9-1.8 V for battery systems. Minimum feature size for DRAM is 0.18 microns. Design for low power will drive development of new CAD tools.

2000-2005: DRAM reaches 4 Gb per chip; supply voltages are 1.5 V for desktop and 0.9 V for battery systems. CMOS devices will be scaled to sub-0.1 micron size. Compared to current technology at 0.6 microns, speed will improve by 7 times, density by 20 times, and power per function by better than 10 times [DAVA95].

2005-2010: DRAM reaches 64 Gb per chip; performance-optimized clock speed reaches 1100 MHz; supply voltage levels at 0.9 V for desktop and battery systems. Minimum feature size for DRAM is 0.07 microns.

2010-2020: Voltage reduction may reach a point of diminishing returns, or may be too close to threshold voltage to make further reductions feasible. New techniques in lithography are required to proceed with scaling to even smaller feature sizes.

3.3 FEASIBILITY AND RISK

As described in Section 2.1, continued advances materials and manufacturing processes may soon be limited by a hierarchy of constraints, from practical issues, such as investment in new plant, to hard limits of fundamental physics. The forecast assumes that these are engineering problems that can be resolved, i.e., that there are no show-stoppers. Near-term advances are almost certain to continue, if only because there is progress to be made on so many fronts. Though the timeline shown above may vary as certain issues require more time to solve than expected, the technology levels forecasted for 2010 seem to be realistic goals for the industry. To continue advances in low power electronics beyond this level will probably require significant improvements in every technology area presented herein.

References

[ABID95] Abidi, Asad A., "Low-Power Radio-Frequency IC's for Portable Communications," *Proc. IEEE,* Vol. 83, No. 4, April 1995, pp. 544-569.

[ACMO96] Overview of the ACMOS Group at USC's Information Sciences Institute Website, http://www.isi.edu/amcos/overview.html, accessed November 11, 1996.

[CHAN95a] Chandrakasan, Anantha P., and R. W. Brodersen, "Low Power Digital CMOS Design," Kluwer Academic Publishers, Boston, 1995.

[CHAN95b] Chandrakasan, Anantha P., and R. W. Brodersen, "Minimizing Power Consumption in Digital CMOS Circuits," *Proc. IEEE,* Vol. 83, No. 4, April 1995, pp. 498-522.

[CHAN96] Chandrakasan, Anantha P., et al., "Design Considerations and Tools for Low-Voltage Digital System Design," presented at the 33rd ACM Design Automation Conference, Las Vegas, NV, June 1996.

[CHOI96] Choi, M., "CMOS Op-Amp Topologies for Low Power Applications," sponsored by the Audio/Video Communication Circuit Consortium, WWW at http://www-mtl.mit.edu/MTL/Report94/IC/IC.abs.09.html, accessed on November 11, 1996.

[DAVA95] Davari, Bijan, R. H. Dennard, and G. G. Shahidi, "CMOS Scaling for High Performance and Low Power - The Next Ten Years," *Proc. IEEE*, Vol. 83, No. 4, April 1995, pp. 595-606.

[FOSS95] Fossum, E. R., "Low Power Camera-on-a-Chip Using CMOS Active Pixel Sensor Technology," conference paper presented at the 1995 IEEE Symposium on Low Power Electronics, San Jose, CA, October 9-11, 1995.

[HARR95] Harris, Erik P. et al., "Technology Directions for Portable Computers," *Proc. IEEE*, Vol. 83, No. 4, April 1995, pp. 636-658.

[IMAN96] Iman, Sasan, and M. Pedram, "POSE: Power Optimization and Synthesis Environment," presented at the 33rd ACM Design Automation Conference, Las Vegas, NV, June 1996.

[ITOH95] Itoh, Kiyoo, K. Sasaki, and Y. Nakagome, "Trends in Low-Power RAM Circuit Technologies," *Proc. IEEE*, Vol. 83, No. 4, April 1995, pp. 524-543.

[JOHN95] Johnston, A. H., G. M. Swift, and D. C. Shaw, "Impact of CMOS Scaling on Single-Event Hard Errors in Space Systems," conference paper presented at the 1995 IEEE Symposium on Low Power Electronics, San Jose, CA, October 9-11, 1995.

[JONE89] Jones, Ross M., "Think Small - In Large Numbers," *Aerospace America*, October 1989, pp. 14-17.

[LEMN94] Lemnios, Zachary J., and K. J. Gabriel, "Low-Power Electronics," *IEEE Journal on Design and Test of Computers*, Vol 11, Winter 1994, pp. 8-13.

[LEMN96a] Lemnios, Zachary J., "Low Power Electronics: A Technology Revolution," presented at the DARPA Low Power Electronics Symposium, on the Web at http://esto.sysplan.com/ESTO/LPE/Presentation/index.html, accessed on September 13, 1996.

[LEMN96b] Lemnios, Zachary J., "Manufacturing Technology Challenges for Low Power Electronics (LPE)," on the Web at http://esto.sysplan.com/ETO/Articles/LPE/Article2.html, accessed on September 13, 1996.

[LIDS96] Lidsky, David, and J. M. Rabaey, "Early Power Exploration - A World Wide Web Application," presented at the 33rd ACM Design Automation Conference, Las Vegas, NV, June 1996. [Note: The PowerPlay application is made available on the Web at http://info-pad.eecs.berkeley.edu/PowerPlay.]

[MANG95] Mangione-Smith, B., "Low Power Communications Protocols: Paging and Beyond," presented at the 1995 IEEE Symposium on Low Power Electronics, IEEE Solid State Circuits Council, San Jose, CA, October 9-11, 1995.

[MEIN95] Meindl, James D., "Low Power Microelectronics: Retrospect and Prospect," *Proc. IEEE*, Vol. 83, No. 4, April 1995, pp. 619-635.

[NASA95] "Technology Directions for the 21st Century," National Aeronautics and Space Administration, Office of Space Communications, May 1996, Volume I, page 1-1 ff.

[NTRS94] "The National Technology Roadmap for Semiconductors," sponsored by the Semiconductor Industry Association, 1994, page B-2.

[PEDR96] Pedram, Massoud, "Power Minimization in IC Design: Principles and Applications," *ACM Transactions on Design Automation of Electronic Systems*, Vol. 1 No. 1, January 1996, pp. 3-56.

[PERS97] Product review from *Personal Computing Magazine*, Ziff-Davis Publishing, Vol. 16, No. 1, January 1997.

[POWE95] Powers, Robert A., "Batteries for Low Power Electronics," *Proc. IEEE*, Vol. 83, No. 4, April 1995, pp. 687-693.

[ROBI96a] Robinson, Ernest Y., H. Helvajian, and S. W. Janson, "Big Benefits from Tiny Technologies," *Aerospace America*, October 1996, pp. 38-43.

[ROBI96b] Robinson, Ernest Y., H. Helvajian, and S. W. Janson, "Small and Smaller: The World of MNT," *Aerospace America*, September 1996, pp. 26-32.

[TERM95] Terman, Lewis M., and R-H Yan, Guest Editors, "Scanning the Issue," *Proc. IEEE*, Vol. 83, No. 4, April 1995, pp. 495-496.

[SCHM96] Schmidt, Richard, and Bob Sugars, "Low-Power Key to Implement 8-mm Tape Drive," *System Design*, February 1996, accessed on the Web at http://www.eedesign.com/Editorial/1996/SystemDesign9602.html.

[SING95] Singh, Deo, et al., "Power Conscious CAD Tools and Methodologies: A Perspective," *Proc. IEEE*, Vol. 83, No. 4, April 1995, pp. 570-594.

[STOR95] Stork, Johannes M. C., "Technology Leverage for Ultra-Low Power Information Systems," *Proc. IEEE*, Vol. 83, No. 4, April 1995, pp. 607-618.

[WOLF96] Wolfe, Andrew, "Opportunities and Obstacles in Low-Power System-Level CAD," presented at the 33rd ACM Design Automation Conference, Las Vegas, NV, June 1996.

CHAPTER 3. TRENDS IN NANOTECHNOLOGY

SUMMARY

Nanotechnology is the deliberate manipulation of individual atoms or molecules by a man-made machine or tool of comparable dimensions for the purpose of constructing desired materials or devices. Efforts to develop such a capability are currently at a very rudimentary level, so this technology has a very high risk associated with it. However, the potential rewards are almost limitless. Applications would revolutionize all aspects of our lives encompassing medicine, computers, agriculture, transportation, manufacturing, and the environment to mention only a few. Nanotechnology eventually may even allow reversal of the aging process, thereby achieving the goal of early explorers seeking the fountain of youth. Since applications are limited only by one's imagination, some proponents of nanotechnology exhibit an enthusiasm that seems excessive to more cautious individuals.

Surely the ability to build custom-made materials on demand from their constituent atoms or molecules – a bottom-up approach to manufacturing – must be regarded as a long-term goal. The first true nanoscale construction is likely to result from the top-down approach typified by the steady progress in miniaturization achieved by microelectronics. Hybrid devices where true nanoscale components are integrated on microscale silicon chips will likely provide the first useful products while proving the feasibility of nanotechnology concepts.

Nanotechnology research is currently at a very early stage. Rudimentary molecular manipulation capabilities have been demonstrated, but they are a long way from practical implementation. Since nanoscale robots or assemblers do not yet exist, much current work is being done in modeling and simulation of nanoscale phenomena to design optimum devices and assembly techniques that can be implemented once such robots or assemblers are developed. Experimental work will then need to be done to refine the designs to overcome engineering difficulties that surely will be encountered.

Potential benefits to NASA of nanotechnology lie in the areas of computing, materials, and spacecraft design. Since computers would be smaller and much more powerful, and materials much stronger and lighter, spacecraft would be more capable and able to carry payloads representing a larger proportion of their total weight. Furthermore, all these benefits would be realized at lower cost than is possible today. Longer missions, including deep-space exploration, would be possible since necessary supplies could be manufactured onboard the spacecraft using locally acquired materials or recycled waste products.

The development of nanotechnology presents exciting possibilities. The risks are great, but the potential benefits are almost unlimited. Successful development of a widespread molecular manufacturing capability undoubtedly would enable NASA Administrator Daniel Goldin's vision of "Faster, Better, Cheaper" to be realized.

1. INTRODUCTION

Nanotechnology is a broad term that has come to mean different things to different people. It is sometimes applied to traditional miniaturization efforts as scale size continues to decrease toward the micron range, but it is more correctly used in referring to manipulation carried out on an atomic scale. Nobel Laureate Richard Feynman, in a famous talk [FEYN59], stated his belief in the feasibility of "manipulating and controlling things on a small scale," thereby giving rise to the field that has come to be known as nanotechnology. K. Eric Drexler, perhaps the leading proponent of nanotechnology, and his followers define it to be the building of machines and other objects on an atomic level by manipulating individual atoms and molecules. The prefix nano comes from nanometer, a unit of measure equal to one billionth (10^{-9}) of a meter – approximately ten times the diameter of a single atom.

The term nanotechnology is also sometimes associated with a particular application area of the technology. Applications include molecular manufacturing, molecular computing, construction of micromachines or micro robots, biology, medicine and genetics, materials science, and others. In this paper, nanotechnology will be interpreted broadly to encompass all such applications where operations actually occur through the deliberate manipulation of individual atoms or molecules by a man-made machine or tool of comparable dimensions. This definition is intended to distinguish nanotechnology from natural processes that have been ongoing in nature for eons, such as ordinary chemical reactions and the assembly of living organisms in accordance with instructions carried in the genetic code, that also occur at the atomic level and involve the arrangement or rearrangement of atoms and molecules.

Please note that in discussing the research reported in this paper, the term nanotechnology sometimes will be used when the work is actually being done at micron scale because of a misuse of the terminology by the workers involved; this misuse of terms often extends to the titles of papers and articles concerning their work. Until standard use of terms is widespread, some confusion is inevitable but usually can be eliminated by noting the dimensional scale of the work being described.

Key features needed for nanotechnology to be useful include the ability to place each atom in the correct location, being able to build essentially any structure that can be specified in atomic detail and is stable according to the laws of chemistry and physics, and achieving low manufacturing costs that do not greatly exceed the cost of the raw materials and energy needed to build the desired structure. Related factors for achieving these capabilities are positional control and self-replication. The former is needed to precisely move atoms and molecules to the desired locations, while the latter is needed to keep down costs by having a large number of identical assemblers operating simultaneously. Without a large number of assemblers, it would take much too long to produce macroscopically appreciable quantities of the items being manufactured.

A useful description of three areas of activity within nanotechnology is provided by Nanothinc at their Web site [NANO96] and is reproduced below.

> "1. Enabling and metaphorical technologies. Enabling technologies make the big smaller, working from the "top-down" to create ever smaller products. Metaphorical technologies could be described as "nanotech large," functioning in very much the same fashion as molecular assembly, but on a larger-than-nano (micro or macro) scale.

2. Nanotechnology tools. These include scientific equipment, devices, and software programs which allow scientists not only to visualize atomic structure, but to manipulate individual atoms, and to model novel molecules.

3. Mature molecular nanotechnology or molecular manufacturing. The use of complex molecular machines (assemblers) capable of reproducing themselves in large numbers and then gathering and positioning other atoms and molecules in desired constructions..."

2. LONG-RANGE TECHNOLOGY TRENDS

2.1 ANALYSIS OF TRENDS

2.1.1 Brief History Of Nanotechnology

Nanotechnology did not really start to materialize until the early- to mid-1980s. Prior to then, chemists and biologists were focused on studying and manipulating molecular structures from the bottom up, while physicists and engineers were concerned with making components smaller and smaller using solid state devices, and ultimately integrated circuit technology, in a top-down approach. Both techniques yielded spectacular results, ranging from new medicines to synthetic fibers to micron-scale computing machines. While the two approaches are complementary to a degree, proponents of each are still sharply divided. For example, Drexler advocates the construction of nanoscale robots or assemblers, capable of self-replication, that can then be programmed to build essentially anything one could desire, from bearings and gears to proteins and computers consisting of atomic scale logic elements. Edward Wolf, with Cornell University's National Nanofabrication Facility, believes in the top-down approach of building small objects, then very small objects, and proceeding until the atomic level is reached. Top-down proponents feel that their approach is more practical and will yield useful results along the way, and tend to be strongly opposed to the bottom-up approach. Indeed, speaking of Drexler's approach, Wolf has said, "The fact is, we have no idea how to do 3-D assembly. I don't want to demean anybody, but nanotechnology has been oversold, in the sense that it's been popularized. Let me put it this way. As scientists, we all have great imaginations. But most of us are a little more cautious about speaking out" [PLAT94].

A key development for the design of atomic scale systems was the invention of the scanning tunneling microscope (STM) in 1981. By means of a stylus hovering over the surface of a material, the STM forms a direct image of the individual atoms constituting the surface. In the same year, Drexler wrote an article in which he claimed the feasibility of human-engineered molecular machines [DREX81]. In 1985, a new form of carbon molecule, dubbed buckminsterfullerene in honor of R. Buckminster Fuller, was discovered. The structures, nicknamed buckyballs, consist of 60 interconnected carbon atoms in the shape of a soccer ball and have been used in a number of nanotechnological efforts as the basis of three-dimensional structures (see, for example, [GLOB96]). In 1989, the first international conference on nanotechnology was held, sponsored by the Foresight Institute founded by Drexler. This was followed in 1990 by the establishment of a new technical journal, called *Nanotechnology*, published by the English Institute of Physics. In the same year, nanotechnology burst on the public scene thanks to the widely disseminated STM image produced by IBM showing the letters I, B, and M formed by arranging 35 xenon atoms on the surface of a nickel crystal. In 1992, Drexler wrote a technical treatise, *Nanosystems*, setting forth the theoretical basis and methods for constructing nanoscale machines and using them for molecular manufacturing [DREX92]. While the assemblers Drexler speaks of cannot be made today, he believes they will be feasible within the next few decades.

Following these early efforts, nanotechnology has picked up momentum, with many companies and universities instituting research efforts in various aspects of the field. Over the past few years, numerous conferences have been devoted to nanotechnology or featured sessions dealing with this area. The Foresight Institute has been sponsoring an annual Conference on Molecular Nanotechnology. The fifth in the series is scheduled for 1997. Another recent conference on biological approaches and novel applications for molecular nanotechnology contained sessions addressing: (1) tools and techniques for molecular nanotechnology, (2) programmable self-assembly systems, (3) nanosensors and detection of biomolecular interactions, and (4) protein and organic-based self-assembly systems. Speakers represented a range of government laboratories, universities, industrial laboratories, and commercial firms.

Another indication of the establishment of nanotechnology as an area of active research and development is the growing number of books devoted to this field. A casual search revealed about 30 books in print dealing with the subject of nanotechnology, ranging from scientific treatises and conference proceedings to popular accounts. A good historical account and survey of current research and applications is contained in the book edited by Crandall [CRAN96]. Regis gives an interesting journalistic account of the field, featuring the personalities of the key players [REGI95], while a more speculative discussion of the implications of nanotechnology for achieving immortality is provided by DuCharme [DUCH95]. Because of the rapid advances occurring in this young field, perhaps one of the best sources of current information on nanotechnology is the World Wide Web. A good starting point containing useful information and references to other Web sites is the home page maintained by Ralph Merkle at Xerox's Palo Alto Research Center, whose URL is *http://nano.xerox.com/nano*. Another useful site is maintained by The Foresight Institute at *http://www.foresight.org*.

2.1.2 U.S. Organizations Active In Nanotechnology

A large number of organizations are now involved in nanotechnology, including research and development laboratories, universities, and commercial firms. Two organizations promoting nanotechnology are The Foresight Institute and Nanothinc. The Foresight Institute, a non-profit organization, sponsors annual nanotechnology conferences and educates both the general public and the research community about the benefits of nanotechnology. Nanothinc is more commercially oriented, providing Web-based services for companies to market their nanotechnology-related products and services. Its information services extend to related enabling technologies, such as supramolecular chemistry, protein engineering, molecular design and modeling software, and such top-down approaches as nanolithography and micromachines. The Institute for Molecular Manufacturing is a non-profit foundation carrying out research directed toward molecular manufacturing. Molecular Manufacturing Enterprises, Inc. provides seed capital and other support to organizations developing key advances in the field of nanotechnology.

Many universities are conducting research relevant to nanotechnology. Some of these will be identified below when specific work being done is discussed. However, a number of nanotechnology research centers have been designated by the National Science Foundation (NSF). They are:

- University of Illinois
- Carnegie Mellon University
- University of Colorado
- University of Maryland
- Mississippi State University
- University of Wisconsin

Some other universities placing particular emphasis on nanotechnology or hosting prominent researchers are:

- Cornell University
- Princeton University
- Rice University
- Rutgers University
- New York University
- University of Southern California, Laboratory for Molecular Robotics
- California Institute of Technology
- University of North Carolina, The Nanomanipulator Project
- MIT, Nanostructures Laboratory
- Stanford University, NanoFabrication Facility
- SUNY Stony Brook, Molecular Structure Laboratory
- Virginia Commonwealth University, Nanosystems and Nanotechnology Laboratory
- Penn State University, NanoNet nanofabrication facility
- Purdue University, Nanoscale physics
- University of Michigan, Michigan Molecular Institute
- Northwestern University

Commercial laboratories involved in nanotechnology include:

- Xerox Palo Alto Research Center
- IBM
- Texas Instruments Nanotechnology Center

Government laboratories and organizations supporting nanotechnology include:

- Advanced Research Projects Agency
- NASA Ames Research Center
- Oak Ridge National Laboratory
- Naval Research Laboratory

2.1.3 Foreign Nanotechnology Efforts [LANG96]

Nanotechnology research is also being pursued in other countries, most notably Japan. Below is a brief description of the research enterprises in Japan, the United Kingdom, Canada, and Australia. In addition to these countries, France, Germany, Sweden, Denmark, Russia, and China also have strong research efforts [NELS95].

The Japanese government has pledged over $200 million to support Japanese nanotechnology research. This funding is awarded through the Ministry of International Trade and Industry (MITI). They hope to have commercial firms supply the needed equipment so that research can be carried out that will enable Japan to become dominant in the manufacture of nanoscale electronic and sensing devices. Two nanotechnology centers will be established: one at Tsukuba Science City emphasizing technology, and one at the University of Tokyo emphasizing science. Japanese companies are expected to form alliances with these centers, contributing additional funding over a ten-year period. MITI and the Research Development Corporation of Japan have also funded six other nanotechnology projects at a level of about $75 million. There is some concern among Japanese researchers, however, that MITI does not understand how to conduct fundamental research and may mismanage their efforts. Also, to date, there has been almost no cross-fertilization between the nanotechnology and biotechnology communities in Japan, which will hamper the development and application of nanoscale technology in biology and medicine.

The nanotechnology research effort in the United Kingdom is divided into two disjoint areas: nanostructures and nanoelectronics. As a result, interdisciplinary efforts common in the United States, such as research on semiconductors coupled with biological nanostructures, are hindered. The National Physical Laboratory coordinates a program funded by the Department of Trade and Industry at about $9 million over four years. A major grant from the Science and Engineering Research Council funds research at the University of Glasgow's Nanoelectronics Research Centre, one of the leading nanotechnology research institutions in the world.

The focus for Canadian nanotechnology research is the Institute for Microstructural Sciences within the National Research Council. Their major nanotechnology effort, begun in 1990, is the Semiconductor Nanostructure Project. This project has focused on identifying nanoscale physical phenomena and device functionality. Participants include the University of Ottawa, Carlton University, and Queen's University, as well as several telecommunications firms.

Because of a shortage of equipment, Australia's work in nanotechnology is limited. Related research is being carried out at a variety of universities and commercial laboratories, however. A strength of Australian research and development, which should prove valuable for work in nanotechnology, is the frequency of cross-disciplinary programs – a result of the relatively small scale of the Australian research enterprise which did not foster the specialization and isolation of disciplines common in larger countries. A study commissioned by the Australian Minister for Science concluded that a major fabrication facility is needed to enable the country to translate scientific research results into commercial products. Funding for such a facility has not yet been provided.

2.2 FUTURE APPLICATIONS

This section examines the potential of nanotechnology for enabling the assembly of materials on an atom-by-atom basis, and possible applications of that ability in a number of different fields.

2.2.1 Manufacturing

In a sense, manufacturing is synonymous with nanotechnology, which is often referred to as molecular manufacturing. As a result, all applications of nanotechnology could be grouped under this heading. Rather than doing so, this section presents a discussion of the potential of nanotechnology to realize such a manufacturing capability.

The RAND Corporation has prepared a report on the application of nanotechnology to molecular manufacturing [NELS95]. One of their findings is that much research is needed just to assess the long-term feasibility of many of the potential benefits attributed to nanotechnology. Proponents claim that efficient manufacturing of structures, devices, and smart products will occur in the relatively near future, that is, within 10 to 30 years. The goal is to produce complex products on demand by assembling molecular components. It is not at all clear, however, whether such an advanced capability can be achieved in the foreseeable future, and if so, how to develop it. The authors claim that the top-down approach of using macroscale components to build nanoscale structures may be beneficial. Achievements using this approach could serve as a proof-of-principle that molecular manufacturing eventually will be feasible. If some means of molecular assembly is not demonstrated within the next decade, the credibility of nanotechnology for revolutionizing manufacturing concepts may be seriously questioned.

Initial products of molecular manufacturing are likely to be sensors, biomedical products, computing and storage devices, and tailored materials for aerospace or other demanding applications. The current development of microelectromechanical systems may point the way for incorporation of nanoscale components into systems with wide applicability, such as automotive parts.

The potential for a commercial micromanufacturing capability is huge, leading to extreme miniaturization in space systems, enhancements in human performance and medical treatment, as well as the manufacture on demand of a broad range of complex products. Since it is not known what will succeed, multiple research paths should be pursued at both the basic and applied levels. These could include self-assembly methods, fullerenes and other novel chemical structures, semiconductor-organic interfaces, parallel use of massive numbers of scanning force microscopes, and use of molecular modeling tools. Modeling and theoretical understanding of molecular manipulation need further development. Also important are demonstration of assembly techniques and the creation and integration of practical components.

2.2.2 Medicine

Applications of nanotechnology to medicine and biology illustrate some of the most dramatic and far-reaching potential consequences for this area of endeavor. Predictions range from the relatively mundane development of sensors and other diagnostic aids all the way to excited claims of immortality through reversing the aging process. There is, in fact, an entire book devoted to the latter subject [DUCH95].

Even on a more modest scale, the potential applications of nanotechnology to medicine are important. Wars against viruses and bacteria could be fought at the molecular level by interfering with their chemical processes or by disassembling them. Miniature submarines carrying specifically tailored drugs could cruise through the bloodstream, guided by nanoscale computers, seeking and destroying offending matter and organisms. As the technology becomes more advanced, nanomachines would be able to repair breaks and cross-links within a cell's deoxyribonucleic acid (DNA). The action of nanomachines, at least initially, would resemble the way the body now fights disease, with antibodies and killer T-cells. However, nanomachines could be much more precise in their actions, doing their job with greater efficiency than natural processes or medicines, and with no side effects. Tumors could be eliminated and nanoscale assemblers used to repair tissue damage. In repairing brain damage, however, while new cells could be grown, information stored in the original cells probably would not be recoverable.

Nanomedicine would allow genetic diseases to be cured. Once the responsible genetic flaw is identified, nanomachines could go into individual cells and edit chromosomes to repair the defect. By launching huge numbers of these machines into the body, the genetic material in all the cells could be corrected in a short time. It is even conceivable that aging could be reversed, if it results from an accumulation of errors in the chromosomes.

Nanotechnology would also be useful in the diagnosis of disease. Current diagnostic procedures, including X-rays, blood tests, and exploratory surgery are invasive or pose some risk to the patient. Also many diseases produce similar symptoms which are often the products of the body's reaction to the offending organism or substance. Through nanotechnology, diseases could be identified at the molecular level.

2.2.3 Computing

Today's computer technology depends on etching patterns on silicon chips. While photolithography is approaching micron and sub-micron scale, nanotechnology holds the promise of much smaller computing elements. In the first nanocomputers, switching will be done mechanically using tiny atomic scale logic rods made of carbon atom chains. It is estimated that a simple processor with about 10^4 switching elements would occupy a volume of about $2\text{x}10^{-4}$ cubic microns [HALL96]. Later designs would employ electronic switching and even "wireless" switching using quantum effects.

Partly as a result of reduced size, the processing capacity and speed of computers will be greatly increased through the use of nanotechnology. Drexler claims that processing power of 10^{18} million instructions per second will be available in desktop computers [DREX92]. Computer memory and storage densities will also become much higher. In combination, the result will be extremely powerful, small computers produced at relatively low cost compared to current hardware produced using micron scale technology.

2.2.4 Space Applications

Nanotechnology has many applications to spacecraft design and space exploration. The National Space Society, which advocates space exploration and development, recently wrote a position paper supporting the pursuit of nanotechnology for space applications [TOTH95]. They claim that the bottom-up approach to nanotechnology could result in space hardware having much better performance and reliability than is available at present, and these benefits could be achieved at a much lower cost. Self-replicating assemblers, operating at the atomic level, have great potential for space exploration and development, as do microelectronics and materials science through reduced payload weight and increased reliability.

Near term benefits of nanotechnology include greatly reducing the size and mass of sensors and actuators, planetary probes, and other spacecraft components. Later, once bottom-up manufacturing capability is achieved, bulk structures for spacecraft components with extremely high strength will be produced. Theoretical strength-to-density ratios 75 times greater than those achieved with modern day alloys are possible. These stronger materials will improve reliability while reducing spacecraft weight to permit bigger payloads and higher orbits. In the electronics area, devices will shrink by three orders of magnitude and occupy three dimensions rather than two. This would allow construction of extremely small, inexpensive inertial guidance systems for use in unmanned exploratory spacecraft, interplanetary probes, and planetary rovers. A network of distributed, embedded devices could be placed on the skin of a spacecraft to serve as sensors or an antenna. The manufacturing capability would also be useful for constructing from available extraterrestrial resources additional items needed in space.

Long-term benefits will result from the ability to bootstrap production through use of self-replicating universal assemblers. Manufacturing costs would be lowered by orders of magnitude, down to about $1 per kilogram. Building tapered tethers extending from the ground to geosynchronous orbit could provide inexpensive access to space. Mature nanomanufacturing systems might make possible affordable closed environment life-support systems capable of sustaining human life using whatever materials were available. Enhanced medical capabilities could enable in-vivo repair of cell damage caused by ionizing cosmic radiation, thereby enabling extended, manned extraterrestrial missions.

2.2.5 Environmental Clean-up

Nanotechnology has the potential to end environmental pollution. The need to cut down trees or mine metals could be greatly lessened or eliminated if large-scale molecular manufacturing, enabling the construction of any desired material from its constituent atoms, becomes economically feasible. The manufacturing process would use clean energy and emit only harmless by-products such as oxygen and water. The ability to recycle garbage on a molecular basis could provide materials needed to build new products. Nanomachines would also be able to reclaim polluted land and water by seeking out toxic chemicals and converting them to harmless compounds. Pesticides and fertilizers used in farming could be made obsolete by nanoscale machines placed inside plants to generate whatever nutrients are needed while attacking any invading microorganisms or harmful insects by disabling their normal cellular processes.

2.2.6 Other Applications

Many more applications of nanotechnology are possible. The possibilities are limited only by one's imagination. Entire books are devoted to speculation about what the world will be like when molecular manufacturing is widespread. A good account, divided roughly into applications within living bodies and external to those bodies, is contained in the book edited by Crandall [CRAN96]. These applications range from merely changing hair or skin color to imagining the existence of a so-called utility fog, millions of nanoscale robots dispersed throughout the air capable of invisibly moving or restraining objects, constructing or disassembling objects, and performing other magic-like functions.

While the goals of nanotechnology are dismissed by some as daydreaming and wishful thinking, achieving the capability to manufacture objects on an atom-by-atom basis surely would lead to a large number of useful applications.

2.3 EMERGING TECHNOLOGIES

Because of the relatively rudimentary state of nanotechnology, there are many research paths, and results are sometimes hard to categorize. Therefore, the following discussion of results will be a synopsis of several pertinent research efforts rather than an integrated treatment of specific topics to which incremental knowledge is being added.

Intelligent Micromachines Produced By Sandia Laboratories [SING96]

Sandia Labs has successfully fabricated an intelligent micromachine consisting of a tiny motor together with an integrated circuit controller, all on a single silicon chip. The fabrication process uses tiny trenches etched on the chip within which the motors are built using microelectronic fabrication techniques. These motors are then surrounded by liquefied silicon dioxide that hardens around them, creating a smooth surface on which the required circuitry is fabricated by photolithography. Removal of the silicon dioxide then frees the motors, enabling mechanical movement of their parts. The trenches used are about 6 microns deep and each resulting micromachine (motor plus controller) is about 1 millimeter square. Two possible applications are tiny drug-delivery devices and inexpensive, long-lasting gyroscopes for military and civilian uses. Transfer of the technology to industrial partners for large-scale production is taking place. The inexpensive manufacturing process could be used to produce tens of thousands of micromachines per day.

Use Of Laser Beams To Manipulate Particles [BAIN96]

Two teams of researchers in the United Kingdom and Australia have demonstrated the use of laser beams to spin a molecule and also to speed up or slow down the spin, and even stop it and reverse its direction. The researchers feel that the technique has potential for use in nanotechnology and, because of its low power, can be used to safely manipulate living tissue.

Creation Of Stable Metallic Cluster Arrays [ANDR96]

Planar arrays of small metal islands (particles) separated by tunnel barriers are of interest for developing nanoscale electronics. Conductivity of the array can be varied by controlling the size of the islands and the strength of the coupling between them. Researchers at Purdue University have been able to create a self-assembled superlattice of gold nanocrystals having mean diameter of 3.7 nanometers. The four-step process used involves synthesis of ultrafine crystals, use of a surfactant coating to provide stability during manipulation, formation of a monolayer film of particles on a solid substrate, and finally, displacement of the surfactant by molecular interconnects creating covalent bonds between adjacent gold particles. Besides adding stability, these bonds serve as "molecular wires" providing controlled electrical coupling between the adjacent islands of gold particles. This cluster network of particles exhibits a nonlinear current-voltage relationship.

Using DNA To Construct Three-Dimensional Objects [UNIS96, BROW96]

Nadrian Seeman at New York University received the 1995 Feynman Prize in Nanotechnology for his work in developing ways to create three-dimensional objects, including cubes and other polyhedra, by using synthetic DNA as a scaffolding on which to construct the desired objects. By manipulating the bases that comprise the DNA, Seeman is able to use it to create branched molecules having desired properties. He has also been able to construct knots and linked-ring molecules. Seeman's work provides a basis for the construction of complex devices on a nanometer scale.

Related work is being done at Northwestern University under the leadership of Chad Mirkin. His group has demonstrated a technique for synthesis of materials by bonding strands of DNA to inorganic nanoparticles. The method allows arrays of particles to be built whose chemical composition, periodic structure, and bond strength can be controlled. The DNA-mediated bonding will allow the creation of novel three-dimensional structures as well as the assembly of two-dimensional structures on a surface. The technique can be used for a range of inorganic substances including metals, semiconductors, and magnetic materials. Besides its use for synthesis, it will aid DNA chemistry studies because the inorganic particles show up on transmission electron microscope images much better than does DNA alone. Mirkin is optimistic that a practical process for synthesis will follow from the experimental demonstrations achieved by his group because of the detailed knowledge now available about DNA processes – knowledge gained from genetic engineering research and practical experience in the biotechnology industry.

Molecular Biology Computation [ROBI96]

Efforts are underway to apply molecular biology to the area of computation. Dr. James Hickman of Science Applications International Corporation and Dr. David Stenger of the Naval Research Laboratory are working on the development of functional neuronal circuits arranged on a patterned, self-assembled substrate. They are now constructing two-cell circuits from living neurons and studying their interaction. The hope is to be able eventually to assemble a biological system and combine it with a solid state system. Attempts are being made to communicate between biological and solid state circuits. At this time, however, it is not even well understood how cells exchange information with each other. Hickman indicates that with adequate funding (about $1 million per year) neurons could be made to self-assemble and grow, making desired connections for communications and computation. A proof of concept bioelectronics neuronal system could be developed within five years, he believes.

Computational Nanotechnology [MERK91]

While the ultimate goal of nanotechnology is the molecular fabrication of machines and devices, present capabilities only permit construction of the most rudimentary molecular structures. However, achievement of such a capability can be aided through the use of computer-assisted modeling and design, which are feasible with current technology. Through the use of computer-aided design (CAD) software, alternatives for molecular manufacturing can be evaluated to select the most promising approach. This should result in a time savings once nanoscale assemblers capable of moving single atoms and molecules are available. For instance, the best materials and structure for a bearing can be determined using computational chemistry tools. When available, the assembler can then be used to construct the bearing, avoiding trial and error manipulation of different materials and structures.

Two modeling techniques are particularly useful: molecular mechanics and *ab initio* methods. Molecular mechanics allows modeling of the positions and trajectories of atomic nuclei without an excessive computational burden. It is able to do this by using energy minimization methods based on the existence of empirically determined potential energy functions, rather than dealing with the quantum mechanical relations that would be needed to model electron clouds. As long as the structures under consideration are relatively rigid without unconstrained torsion, modeling the forces between adjacent atomic nuclei treated as point masses is sufficient. Current tools, running on a personal computer, permit such energy minimizations to be performed on systems consisting of hundreds of thousands of atoms, or more. This simplified modeling could, of course, introduce errors or imprecision in the calculations. However, if the modeling errors are small with respect to the scale of errors that will produce incorrect functioning of the modeled device, then model results will tend to be reliable. Since the modeler is selecting the structures to be built, he can deliberately restrict attention to those that are robust with respect to the computational tools he is using.

If the structures of interest are not sufficiently rigid, or if actual chemical reactions must be modeled, molecular mechanics models will not be adequate. In that event, more complicated, higher-order *ab initio* techniques must be used that severely restrict the number of atoms that can be modeled. For example, perhaps only a couple of dozen heavy atoms may constitute a computationally tractable limit. This should be sufficient to analyze the removal or addition of a small number of atoms from a specific site on a molecular structure. Synthesis of a larger structure could then be performed by repeated application of the computational tools to specific local sites.

Not every molecular machine that one might wish to construct fits the constraints of the foregoing modeling methods, but many do, including bearings, computers, and robotic arms. By restricting attention to these devices, currently available modeling approaches and methods are adequate to perform preliminary design and tradeoff studies that will speed up the eventual fabrication of the devices.

Nanoscale Techniques Useful For Biotechnology [EDGI94, WHIT95]

There are a number of areas within biotechnology that could benefit from nanoscale technology. Recent developments, though really on a microscale rather than nanoscale, illustrate the potential of this new technology. Robert Austin of Princeton University developed an array of micron-sized posts on a silicon chip to serve as a matrix for DNA electrophoresis. The uniformity, known structure, and two-dimensional character of the matrix is an improvement over the type of gels that had been in use for electrophoresis. The resulting increase in speed of DNA fractionation should be a big benefit to genetics research, such as the human genome project. Further applications are also foreseen, such as coupling the technique with optical or electrical detection to enable the analysis of single cells for diagnostic and therapeutic use. The technique's portability and reproducibility should make it useful also in health care and environmental detection applications.

Harvey Hoch of Cornell University discovered that spatial features can be significant in the infection of plants by rust fungi. By fabricating submicron height ridges similar to those surrounding a vulnerable area on a plant's leaf, he found that the fungi generated infection structures in response to ridges whose height was between 0.4 and 0.7 microns. No response was produced for ridges less than 0.2 or greater than 1 micron high. These findings could lead to modification of commercial crops to make them resistant to the rust fungi by producing ridges of optimal height. In related work, Chris Wilkinson at the University of Glasgow found that severed tendons will join properly if grown on a quartz surface with parallel microgrooves machined on its surface. His group developed a procedure to emboss single-cell wide grooves on the surface of biodegradable materials with the hope of using the resulting material as an *in vivo* splint to allow severed tendons to repair themselves.

George Whitesides at Harvard University synthesized a single biologically active layer on a gold film substrate and then developed several ways to apply submicron-scale patterns to the layer. The initial biological application of these techniques is to study the relationship between the shape of cells and their function. For example, using a cell-size square pattern, he was able to get cells to assume a rectangular shape. Next, the effects of this shape on DNA synthesis, the control of growth and differentiation, and protein production will be studied.

Survey Of Recent Achievements In Nanoelectronics [MITR96]

The MITRE Corporation has compiled a list of research developments they believe are significant breakthroughs on the way to developing practical nanoelectronic devices and computers. A synopsis of several of these efforts follows.

- Researchers at the University of South Carolina and Penn State University have demonstrated a nanoscale wire consisting of a single chain molecule capable of conducting electricity from a gold lead, to which it is attached at one end, to the tip of an STM probe. Such structures have been proposed previously, but this is the first actually shown to conduct electricity.

- Lent and Porod at the University of Notre Dame have used quantum-mechanical modeling and simulation to show the feasibility of sending a signal along a line of quantum-dot cells, capable of switching between two states, without any current flowing. Efforts are underway to fabricate wireless electronic logic structures based on these quantum-dot cells.

- A group at IBM Almaden Research Laboratory is fabricating and testing so-called quantum corrals, primitive nanoscale devices for the manipulation of charge on the surface of a solid. The enclosures are 2 to 5 nanometers across, formed using a few dozen atoms individually positioned using an STM.

- The Nanoelectronics Group at Texas Instruments Corporation are embedding nanoscale quantum devices on a chip with conventional microelectronic logic components. Such hybrid logic is an important intermediate step to achieve much higher logic density while retaining the features and reliability of conventional microelectronics.

- Researchers at the IBM Zurich Research Laboratory have been able to manipulate single molecules using an STM operating at room temperature. This is of practical significance for large-scale fabrication to avoid the complexity and cost associated with the cryogenic cooling of equipment.

- An array of microelectromechanical STMs has been fabricated on a chip at Cornell University. Related work on micro atomic force microscopes (AFMs) is underway at Stanford University. Such devices should be useful for the mass fabrication of nanostructures through atomic manipulation rather than chemistry-based synthesis.

If such STM and AFM arrays can themselves be mass manufactured, distributed manufacturing of nanostructures and nanodevices may become possible.

Nanotechnology Research At NASA Ames Research Center [GLOB96b]

NASA's Ames Research Center recognizes the promise of molecular nanotechnology for aerospace system design, and because of their strength in parallel supercomputing as well as computational chemistry, has identified computational molecular technology as the area in which they can make the greatest contribution. They hope to lead a nationwide network of laboratories using computation to understand, design, and control programmable molecular machines, their products, and related manufacturing processes.

Three application areas for molecular manufacturing are of interest to NASA: launch vehicle structural materials, computer components, and small-scale spacecraft. The current transportation cost of lifting an object into an orbit several hundred miles above the Earth's surface is about $10,000 per pound. The promise of extremely strong diamondoid materials produced by nanoscale manufacturing would allow 9 to 12 percent of vehicle launch weight to be devoted to payload versus 1 to 5 percent using titanium structures. Diamondoid materials could also be used to produce extremely high density computer memory components (approximately 10^{15} bytes per cm^2) and rods for a nanoscale mechanical central processing unit able to achieve 10^{18} million instructions per second in a highly parallel desktop computer. Finally, the use of atomically precise components should allow a radical decrease in spacecraft size, thereby enabling new missions using large numbers of very small spacecraft.

The NASA Ames Research Center computational nanotechnology initiative is pursuing the following activities:

- Fullerene nanotechnology – At present, the focus is on development of (1) gears based upon carbon nanotubes, (2) design software, (3) parallelized molecular dynamics based on an atomic scale energy potential function, and (4) quantum calculations related to the gear teeth.
- Diamondoid mechanosynthesis – Investigation of reaction pathways for diamond creation.
- Properties of clusters – Investigation of the properties of matter for clusters up to sizes exhibiting bulk properties.
- Nanotube strength – Using potential functions to determine the tensile strength of different nanotube configurations.
- Entropy and temperature effects – Development of methods to address the effects of entropy and finite temperature on nanodevice stability.
- High density memory – Investigation of diamondoid structures for achieving approximately 10^{15} bytes/cm^2 storage density.
- Laplacian of the electronic charge density – Use of modeling software to visualize and understand molecular electronic structure and reactivity.

- Support of a grant program in computational molecular nanotechnology. Proposals were due 15 October 1996, with work to start around December 1996. Suggested research topics include diamondoid mechanosynthesis, self assembly, simulation, molecular CAD, system design, and component design.

In addition to these activities, a number of collaborative efforts are underway with researchers at the California Institute of Technology, Xerox PARC, North Carolina State University, the University of California at Santa Clara, and Hughes Aircraft Company.

It is understood that nanotechnology, while being a high risk endeavor, promises significant benefits for future NASA missions. Consequently, it is regarded as a long-term research effort directed toward the following goals:

- Develop a software environment conducive to research within one year.

- Develop a precise design within five to seven years for an assembler capable of building long thin structures (e.g., fibers) with tensile strength comparable to diamond. Such fibers would be of use in the production of aerospace composites.

- Develop a molecular manufacturing CAD system within ten to fifteen years.

- Develop an efficient molecular manufacturing system design within fifteen to twenty-five years.

These activities will be done with a combination of in-house staff and grant support to academic and industrial research laboratories. Where possible, collaboration also will be established with experimentalists in order to validate numerical results obtained through computational research. Heavy use is expected to be made of NASA's numerical aerospace simulation facility's supercomputer center during all phases of the research.

The ARPA Ultra Dense, Ultra Fast Computing Components (ULTRA) Program [ARPA96]

The following description of this Advanced Research Projects Agency (ARPA) Electronics Technology Office program is taken directly from their ULTRA Web page, located at http://eto.sysplan.com/ ETO/ULTRA/index.html.

> "The goals of the Ultra Dense, Ultra Fast Computing Components/Nanoelectronics program are to explore and develop material, processing technologies, quantum and conventional devices and device architectures for a next generation of information processing systems. The Ultra program seeks improved speed, density, and functionality beyond that achieved by scaling transistors. These improvements should manifest themselves in systems operating 10 to 100 times faster than current systems, and denser by a factor of five to 100.
>
> Ultra Electronics, Phase I
>
> Phase I of the Ultra Electronics program explores, assesses, and benchmarks alternative electronic approaches to embedded and stand-alone computing architectures.

The program has demonstrated methods for applying novel quantum well electronic devices to improve densities of integrated electronic devices and developed methods of improving the control of epitaxial deposition to realize these devices. Other achievements include developing nanoprobes to study nanometer material structures and devices with picosecond time resolution.

Ultra Electronics/Nanoelectronics, Phase II

Phase II of the program further develops the most promising approaches that were identified in Phase I. A thrust in nanoelectronics includes the design, fabrication, and testing of electronic devices with critical feature sizes below 0.1 microns. Combining conventional transistors with nanoelectronic devices will greatly reduce the complexity and size of sophisticated Department of Defense circuits. This approach, when applied to silicon-based nanoelectronics, will allow nanoelectronics to leverage all the continuing improvements of conventional electronic devices, while providing the density improvements of nanoelectronics. Other thrusts include developing silicon-based nanoelectronics, chemical self-assembly techniques for nanoelectronics and improved semiconductor processing, and molecular beam epitaxy (MBE) in-situ process control and other fabrication techniques for nanoelectronics."

3. RESULTS

3.1 IMPACT ON NASA

Realization of the promise of nanotechnology would have profound effects on all areas of life and, consequently, on NASA operations as well. At present, its effects are felt through pressure to contribute to this nascent technology through research. As mentioned above, NASA Ames Research Center is active in the area of computational nanotechnology through both in-house efforts and funded external research. The greatest impact on NASA, however, will be through application of nanotechnology to NASA's missions.

Molecular manufacturing would result in much smaller, cheaper, stronger, and more reliable components. Computers would be much more powerful with greater storage capability while being very small and lightweight. This would be of great value for use onboard spacecraft. The ability to produce materials that are much stronger yet lighter than current materials also promises significant benefits by reducing spacecraft weight while increasing performance. As a result, larger payloads would be possible and higher orbits would be feasible. It currently costs about $10,000 per pound to place a spacecraft into an orbit several hundred miles above the Earth. For example, the space shuttle program costs about $3 billion per year for six to eight launches with 50-60,000 pounds of payload. Communications satellites also are very costly to launch. While use of titanium allows less than five percent of the weight of a single-stage spacecraft (prior to launch) to be devoted to payload, use of hypothetical diamondoid material would permit payload to be from nine to twelve percent of launch weight [MCKE95].

Extended space missions at greater distances from Earth eventually will be possible. Needed supplies could be restocked by manufacturing whatever is needed in space using available materials. This capability would have enormous implications for both manned and unmanned missions of the future. The lowering of costs will enable NASA to sustain more simultaneous missions or to decrease the interval between successive missions.

Not only onboard applications would benefit from reductions in equipment weight and power requirements. Planetary exploration would be aided by the availability of miniaturized robots covered by nanoscale sensors to facilitate both negotiating the terrain and manipulating objects encountered on distant planets. Development of nanoscale, three-dimensional electronic components would enable construction of extremely small, inexpensive inertial guidance systems for use in such planetary rovers. The small size of the robots will also ease the burden of launching them on their journey to other planets or, alternatively, permit more of them to be included in the mission. Launch technology would also benefit from the development of very lightweight, high tensile strength materials that could be used for tethered launches into geosynchronous earth orbit.

Besides benefiting onboard processing, reduced size and greatly increased computer processing power made possible by the application of nanotechnology would facilitate data capture and level zero processing on the ground to ensure the integrity and availability of mission data. Nanotechnology also could make possible very high density storage media to accommodate the huge amounts of data resulting from various programs such as the Hubble Space Telescope and Earth Observing System. This combined capability to process, store, and disseminate invaluable data downlinked from space will be a great asset to both NASA and the end users of mission information, located worldwide. The availability of several years worth of such data – practicable only through the use of high density mass storage devices – will increase its value to researchers.

3.2 TECHNOLOGY ROADMAP

Because nanotechnology is very much in the research stage, it is extremely difficult to predict the availability of various capabilities with any accuracy. Even the evolution of research capability is hard to define. NASA hopes to have an assembler design worked out in five to seven years, and a validated, efficient molecular manufacturing system design ready within fifteen to twenty-five years [GLOB96b].

Extrapolating current trends in microelectronic miniaturization, using Moore's Law, indicates that nanoscale electronics will be available sometime between 2010 and 2020. If a hybrid approach is adopted, with nanoscale quantum effect devices embedded in microelectronic chips, this form of nanoelectronics might be available as early as 2005. Fully integrated nanocomputers probably would not be a reality until after 2020.

Since extrapolation of current trends is risky, another approach to forecasting is to use expert opinion. In August 1995, *Wired* magazine reported estimates for the achievement of several nanotechnology capabilities provided by the following five experts:

- Robert Birge – Distinguished Professor of Chemistry, Syracuse University
- Donald Brenner – Associate Professor of Materials Science, NC State University
- K. Eric Drexler – Chairman, Foresight Institute

- J. Storrs Hall – computer scientist, Rutgers University
- Richard E. Smalley – Professor of Chemistry and Physics, Rice University and 1996 Nobel Laureate in Chemistry

Table 3.1 summarizes their estimates for the year in which the indicated capabilities will be realized. Precise definitions of the capabilities and any qualifications are not known, so the estimates should be taken as only rough guesses. They do, however, indicate the thinking of some of the most prominent workers in nanotechnology.

Table 3.1. Nanotechnology Capability Predictions

Application	Birge	Brenner	Drexler	Hall	Smalley
Molecular assembler	2005	2025	2015	2010	2000
Nanocomputer	2040	2040	2017	2010	2100
Cell repair	2030	2035	2018	2050	2010
Commercial product	2002	2000	2015	2005	2000

While the range of these estimates is rather large, the panel feels that molecular assembly will be a reality by 2025, while a nanocomputer will be available about 2040 (according to all but one of the experts).

3.3 FEASIBILITY AND RISK

Since it is such an ambitious undertaking, the development of nanotechnology is an extremely high risk endeavor. Some of the risks will be discussed below. One real risk of a different sort is failure to invest sufficient money in nanotechnology research. Since other countries, notably Japan, have organized research efforts underway, if the United States does not support an adequate research program, we will fall behind in the economic, military, commercial, and space-related benefits that may result from the successful development of a nanotechnology capability.

There are three promising routes to the development of a molecular assembler: genetic engineering, physical chemistry, and scanning probe microscopy. Since there is uncertainty as to which approach is best, it is probably wise to fund work on all three. There is a good chance that some approaches might not work, however, so the funding devoted to each should not be too great.

There are many people who feel that the whole concept of nanotechnology is pie-in-the-sky. While these individuals are probably being too pessimistic, their objections do point up some of the risks associated with the development of nanotechnology [STIX96]. Keeping atoms in place during assembly may prove to be exceedingly difficult. For example, to produce the atomic IBM logo mentioned earlier, researchers had to work at extremely high vacuum and supercooled temperatures using inert xenon atoms. Atoms of most elements that would be useful for nanotechnology are very mobile and reactive under normal conditions. There is also the question of how the assemblers would know their own orientation and the location of atoms to be moved. Power also must be provided to the assemblers. Proponents of nanotechnology respond by claiming that these and all other objections to the feasibility of nanotechnology are dealt with in Drexler's book *Nanosystems* [DREX92].

It seems that the biggest critics of nanotechnology are those working at the micron level, that is, the proponents of the top-down approach to miniaturization. For instance, Phillip Barth of Hewlett-Packard says that while *Nanosystems* has a plausible argument for everything, it does not provide any detailed answers. In particular, engineering details are not discussed. While Drexler's nanobearings may be molecularly stable, the stability of intermediate structures used in the synthesis of the bearings is not addressed. Although these objections do not prove the impossibility of molecular manufacturing, they do emphasize the very rudimentary state of current understanding. It will probably take a lot longer than the optimistic proponents of nanotechnology acknowledge before practical devices are a reality.

The top-down approach has so far proven itself by continuing to reduce the scale and increase the density of electronic components that can be placed on a silicon chip. It appears capable of producing useful results sooner than the bottom-up approach. Perhaps a hybrid approach would be the best way to reach an early demonstration of the feasibility of molecular scale assembly. One example of such a hybrid device would be the embedding of quantum well structures on a conventional micron scale silicon chip, mentioned earlier. Bottom-up synthesis of devices from atoms and molecules is best regarded as a long-term goal of nanotechnology.

Since molecular scale assemblers are not available, much of the current research in nanotechnology concerns modeling and simulation of the chemistry and physics of atomic structures. This so-called computational nanotechnology is a reasonable undertaking to try to understand the behavior of atomic structures and to design useful, stable devices, but it must be remembered that this is just a theoretical exercise. The validity of the modeling results must eventually be proven in the laboratory using real atoms and molecules under actual environmental conditions. As in any engineering endeavor, many adjustments undoubtedly will need to be made before the idealized design functions in the desired manner. To be sure, insight will be gained during the design iterations, but realization of practical nanoscale assemblers and devices will probably take much longer than anticipated.

The outlook for the development of practical nanotechnology is probably not as bleak as its critics claim, nor as rosy as its proponents would like us to believe. As Richard Feynman [FEYN59] stated in his seminal talk, there does not appear to be anything in the laws of physics that prevents molecular manufacturing. However, it definitely should be regarded as a long-term goal. Even the most rudimentary practical products are probably at least 20 years away.

References

[ANDR96] Andres, R. P. et al., "Self-Assembly of a Two-Dimensional Superlattice of Molecularly Linked Metal Clusters," *Science*, 20 September 1996, p. 1631.

[ARPA96] ULTRA home page, located on the Web at http://eto.sysplan.com/ETO/ULTRA/index.html.

[BAIN96] Bains, S., "Helical Beams Give Particles a Whirl," *Science*, 5 July 1996, p. 36.

[BROW96] Brown, C., "DNA is Used to Build Inorganic Semiconductor Structures," *Electronic Engineering Times*, 25 September 1996.

[CRAN96] Crandall, B. C., Ed., "Nanotechnology: Molecular Speculations on Global Abundance," MIT Press, Cambridge, MA, 1996.

[DREX81] Drexler, K. E., "Molecular Engineering: an approach to the development of general capabilities for molecular manipulation," *Proceedings of the National Academy of Science*, September 1981, pp. 5275-5278.

[DREX92] Drexler, K. E,. "Nanosystems: Molecular Machinery, Manufacturing, and Computing," John Wiley & Sons, New York, 1992.

[DUCH95] DuCharme, W. M., "Becoming Immortal: Nanotechnology, You, and the Demise of Death," Access Publishers Network, 1995.

[EDGI94] Edgington, S. M., "Biotech's New Nanotools," *Biotech*, May 1994, available on the Web at http://www.enews.com/magazines/bio_tech/archive/050194.5.html.

[FEYN59] Feynman, R. P., "There's Plenty of Room at the Bottom," *Journal of Microelectromechanical Systems*, March 1992, also available on the Web at http://nano.xerox.com/nanotech/feynman.html.

[GLOB96a] Globus, A. and R. Jaffe, "NanoDesign: Concepts and Software for a Nanotechnology Based on Functionalized Fullerenes," NASA Ames Research Center, WWW at http://www.nas.nasa.gov/NAS/Projects/nanotechnology/publications/MGMS_EC1/NanoDesign/paper.html, 1996.

[GLOB96b] Globus, A. et al., "Computational Molecular Nanotechnology at NASA Ames Research Center, 1996," WWW at http://www.nas.nasa.gov/NAS/Projects/nanotechnology/publications/MGMS_EC1/program/paper.html, 1996.

[LANG96] Langbein, M., "An Introduction to Nanotechnology," WWW at http://nanothinc.com/NanoWorld/Introduction/Articles/Melae_Langbein/Melae_Langbein.html, 1996.

[MCKE95] McKendree, T., "Implications of Molecular Nanotechnology: Technical Performance Parameters on Previously Defined Space System Architectures," The Fourth Foresight Conference on Molecular Nanotechnology, Palo Alto, CA, November 1995.

[MERK91] Merkle, R. C., "Computational nanotechnology," edited and updated Web version of paper originally appearing in *Nanotechnology*, 1991, p. 134, available at http://nano.xerox.com/nanotech/compNano.html.

[MITR96] Top 10 Recent Achievements in Nanoelectronics, WWW at http://www.mitre.org/research/nanotech/ten_achievements.html, 1996.

[NANO96] Nanothinc Web page, http://www.nanothinc.com/Nanothinc/WhoWeAre/whoweare.html.

[NELS95] Nelson, M. and C. Shipbaugh, "The Potential of Nanotechnology for Molecular Manufacturing," The RAND Corporation, document number MR-615-RC, summary on WWW at http://www.rand.org/publications/MR/MR615/mr615.html, 1995.

[PLAT94] Platt, C., "Nanotech: Engines of Hyperbole?," *Wired Magazine*, December 1994.

[REGI95] Regis, E., "Nano: The Emerging Science of Nanotechnology," Little Brown, 1995.

[ROBI96] Robinson, C. A. Jr., "Molecular Biology Computation Captures International Research," *Signal Magazine*, February 1996.

[SING96] Singer, N., "Sandia team produces intelligent micromachines," Sandia Lab News, March 1996.

[STIX96] Stix, G., "Waiting for Breakthroughs," Scientific American, available on the Web at http://www.sciam.com/WEB/exhibit/040000trends.html, April 1996.

[TOTH95] Toth-Fejel, T. and T. McKendree, "NSS Position Paper on Space and Molecular Nanotechnology," National Space Society, available on the Web at http://www.islandone.org/MMSG/NSSNanoPosition.html, 1995.

[UNIS96] UniSci Archives, "NYU Chemist Nadrian Seeman Wins Feynman Prize," WWW at http://unisci.com/seeman.htm, 1996.

[WHIT95] Whitesides, G., "Self-Assembling Materials," *Scientific American*, September 1995, p. 146.

CHAPTER 4. BROADBAND COMMUNICATIONS SATELLITE SERVICES

SUMMARY

Recent filings for next-generation broadband satellite communications services in the Ka band show that satellite service providers will seek to compete with terrestrial networks for commercial services. It is highly likely that this trend will continue with the generation-after-next satellite systems. Service providers will seek to exploit advantages of the satellite-based architecture, such as, a relatively rapid roll-out time and low incremental last mile cost. For some consumer-oriented services, such as broadcast of entertainment content, satellites are clearly superior, because the footprint can be nation-wide (a high number of "homes passed"). The advantage is somewhat less clear for video-on-demand service, where the aggregate data rate scales in proportion to the amount of video content. However, the more cost-effective near-video-on-demand variants of this service are compatible with a satellite platform, and will likely be pursued for commercial deployment.

The business case for bandwidth-on-demand service is more difficult to make because the satellite system scales in proportion to the number of users, rather than in proportion to the number of programming ("content") channels carried. This means that for a given satellite capacity, the number of simultaneous customers using the service may be much lower for this type of service. However, it is apparent that companies will pursue this type of service as well, and it will undoubtedly be the driver of new technology.

Satellite system architectures that can potentially meet the generation-after-next broadband satellite communications services requirements proposed in Section 3 of this report are examined in Section 4. The philosophy used in Section 4 was to propose selected system architectures and indicate technologies that should be explored to satisfy the cited requirements. Only a high level discussion of the selected technologies is provided, given the assumption that NASA needs to focus on the range of architectures that are deemed relevant to NASA's internal requirements and those of systems that the NASA staff deem worthy of investment, to leverage as a US investment to support future US commercial and DoD interests.

A major technology focus for generation-after-next broadband satellite communications services is on the need to provide more bandwidth, i.e., a focus on Ka-band (30/20 GHz) and the 50/40 GHz band. Given the migration to these frequencies, the need exists for higher efficiency transmitters (both space and terminal), more dynamic switching and connectivity to support user dynamics, more efficient use of existing resources through frequency re-use (via both spatial and polarization diversity), more adaptive bandwidth vs power efficient modulation, forward error correction coding (including Turbo codes and bit/modulation symbol interleaving), and much expanded use of the variable bit rate formats of dynamic multiplexing techniques such as Asynchronous Transfer Mode (ATM) based technologies.

1. INTRODUCTION

The purpose of this report is to identify the services which may be offered by the generation-after-next satellite communication (SATCOM) systems, and the technologies that will be needed to support and implement these services. This report has been prepared to provide information and guidance to the NASA LeRC to assist in planning its future research programs which will provide selected government developed technologies which will enable the deployment of these future SATCOMs. These technologies may well complement those being developed in the commercial world, and, in fact, may even be somewhat riskier then what industry is willing to take on. However, it is believed that the future of the SATCOMs discussed herein is almost exclusively in the hands of commercial enterprise. NASA LeRC perceives its role as a key component technology supplier, with direct relationships to industry in support of its goals.

One important parameter in the discussion of SATCOMs is the earth orbital distance; so, some useful definitions follow. The geosynchronous earth orbit (GEO), or its equivalent, geostationary satellite orbit (GSO) is at a position 35,860 kilometers above the equator, where the period of revolution is 24 hours. A satellite there can maintain a constant elevation and azimuth relative to any point on Earth, and experiences with about a quarter-second latency or delay in round-trip signal time. The low earth orbit (LEO) is in the 400 to 1400 kilometer range. Earth coverage here requires many satellites, however, being about 50 times nearer offers significant power advantages, low latency, and technical performance comparable to terrestrial systems. These two extremes provide a multitude of options and trades in the design and architecture of future SATCOM systems. In fact, some system concepts combine GEOs and LEOs operationally to provide a proposed suite of services. Finally, there is the medium earth orbit (MEO) in the 8000 to 14000 kilometer range, which has similar but diminished advantages compared to the LEO, combined with greater lifetime on orbit due to lower atmospheric drag. There are few MEO systems planned at this time.

For the purposes of this report, current generation SATCOMs are those to be in place in the 1998 timeframe, such as those offering handheld mobile voice (e.g., Iridium and Globalstar), data (e.g., ORBCOMM) and broadcast video services (e.g., Hughes DirecTV; this effectively introduces the era of the "Big LEO" SATCOM systems providing global module satellite telecommunications services. The next generation SATCOMs are those that will offer a variety of global broadband services, and are scheduled to be deployed in the 2002 timeframe (e.g., Teledesic, Celestri, and Spaceway). These are the subject of a flurry of recent Ka band application filings with the Federal Communications Commission (FCC), the great majority of which propose GEO systems. The generation-after-next SATCOMs are those that will be deployed in the 2003-to-2010 timeframe. The characteristics of these have been developed in this report using projections from next generation services and the best estimates of SAIC engineers of the most likely user applications, the driving requirements, and the resulting supporting technologies. There is very little published information available on these systems, so to some degree, these projections define services and operations that mirror plans for broadband terrestrial telecommunications. It is our opinion that the future of SATCOMs is for seamless interoperability between terrestrial and satellite communications. For NASA LeRC to make its intended impact on the implementation of these systems, it must begin now its R&D program planning for the technologies needed beyond 2003.

There are many terms that are used to describe often overlapping technical concepts or that are used differently to describe the same concept. The following definitions are used in this report:

Applications — The end-result of the communication process desired by and available to the end user, such as viewing a movie, authorizing a credit-card purchase, or viewing Internet Web sites.

Services — The communication system's technical implementation to deliver applications to the user, such as direct video broadcast, high-bandwidth data transmission, or bandwidth-on-demand.

Requirements — The technical characteristics or parameters of the communication system needed to provide a service, such as data rate, bit-error rate, latency restrictions, and transmission protocols.

Architecture/design — A specific communication system defined to deliver specified services, arrived at by combining requirements with spectrum/orbit considerations.

Technology — The physical components and processing needed to implement an architecture/design, such as phased array antennas, millimeter wave components, modulation schemes, and signal processing.

A systematic view of how these terms relate is shown in the following flow diagram:

Application
|
|
Service
|
|
Requirements
|
|
Architecture/Design
|
|
Technology

Feedback can be expected between the Requirements, Architecture/Design and the supporting Technology.

The focus of this report is on future services and technologies. However, there is an intimate relationship among the requirements, the technologies, and the architecture/design, therefore, these must all be discussed together. This report is not intended to produce the specific design for a futuristic SATCOM. It does however, at a high level, discuss the major issues that would be a part of such a design. We have projected the future SATCOM services that we think are the most interesting and challenging; we have focused on the key requirements for these services; we have discussed the general impact of architecture and design on the requirements and the supporting technologies; and, we have identified some of the major supporting technology areas which should be considered for R&D to implement the projected services. This report does not give an exhaustive treatment of any of these issues, and may leave as many open issues as those treated. However, it is a beginning, and may be a template for additional work. We believe that there is a great deal more that can be done to explore the impact of architecture/designs and to identify and discuss technologies.

With regard to available information on these future systems, the FCC applications are a good source of system and service descriptions, up to next generation. However, the FCC does not have this information on line, so it is difficult to access. Also, the FCC likely would not entertain applications for systems not to be deployed for more than seven years, and does not encourage reserving spectrum that far into the future. Finally, even if industry has such plans for systems beyond 2003, and maybe some unique ideas, these would be proprietary, and not likely to be made public this much in advance of deployment. So, for good or bad, we have provided our best engineering estimates of what the future holds.

Finally, we believe that NASA LeRC is currently most interested in technologies associated with power amplifiers, antennas, modulation and coding, and networks. So this report includes these, perhaps to the exclusion of some other important ones. Some of the additional work to be done, as suggested in this report, should address other technology areas that may not currently be represented in the LeRC R&D programs.

2. EMERGING BROADBAND SATELLITE SERVICES

2.1 APPLICATIONS

Applications as defined in the Introduction refer to end uses of electronic media such as the broadcast of television programming, or browsing the World Wide Web. It is necessary to first identify and quantify these applications in order to be able to project what sort of services will be required to transport them. In this subsection we have collated a partial set of commercial applications that are currently being carried by terrestrial and satellite communications systems. These applications were identified through a survey of recent literature, including especially a market survey done for NASA [3-1]. Each application is briefly defined, followed by a Table of selected technical characteristics for each. Key quantitative characteristics include the user data rate, that is, the data rate occupied by the content that actually reaches the end user; the network topology, for example whether it is a broadcast topology or a point-to-point topology; and the latency, or the time that the user must wait before a response to his input occurs. The following definitions apply:

Audio burst – asymmetric high-speed transfer of audio information from a source to a destination.

Audio-on-demand – provision of audio information where the user has real-time control over transmission start time and functions such as fast forward, pause, and rewind via a return transmission path.

Broadcast audio – one-way transmission of audio signals from a single source to a multitude of destinations.

Broadcast video – one-way transmission of video signals from a single source to a multitude of destinations.

Database mirroring – asymmetric transfer of databases from one Internet server to another to provide an alternate site from which files can be downloaded by users.

Distance learning – instruction where the teacher or training materials and students are in different locations. Various scenarios are possible, ranging from an interactive connection between two computers, to multi-site audio conferencing, to multi-site video conferencing involving voice, image, video, and data transmission.

Electronic gaming – symmetric point-to-point connection between computers allowing two opponents to play computer-based games with each other in real time.

E-mail – a store-and-forward electronic message service permitting the exchange of messages between a sender and one or more destinations.

File transfer – asymmetric high-speed transfer of electronic data files from a source to a destination.

High-definition television (HDTV) –. This is a new high-resolution form of television requiring higher transmission bandwidth and providing much better image quality than conventional television.

Information services – asymmetric query and response service where a user requests information, typically at a low data rate, and responses are returned, usually in a high data-rate burst.

Internet file transfer – asymmetric connection allowing requests for and transfer of computer files over the Internet.

LAN-to-LAN connections – two-way high-speed transmission passing data between two local area networks.

Medical imaging – the transfer, usually asymmetric, of imaging information such as X-ray, MRI, and ultrasound images used to permit diagnosis and consultation by specialists located remotely from the patient.

Near video-on-demand (NVOD) – video transmission giving the user some control over the incoming video stream (e.g., start time, fast-forward, pause, and rewind) as in true video-on- demand, but requiring less bandwidth than video-on-demand.

On-line Internet service – various applications such as e-mail, file transfer, and Web browsing, that use the Internet.

Point-of-sale transactions – one-way or two-way transmission of information about a retail purchase (such as credit card verification, inventory control, or shipping information) made at the time of the transaction.

Remote monitoring – one-way data transmission providing status information on devices, appliances, or systems to computers or human attendants at a remote location.

Software distribution – asymmetric connection for the purpose of electronically delivering software from a server to a user.

Video burst – asymmetric high-speed transfer of video information from a source to a destination.

Video-on-demand (VOD) –provision of video information where the user has real-time control over transmission start time and functions such as fast forward, pause, and rewind via a return transmission path.

Video teleconference – two-way television transmission, involving two or more locations, allowing the participants to interact with one another both verbally and visually.

Voice conferencing – simultaneous symmetric voice connection involving three or more locations.

Voice telephony – two-way symmetric audio transmission between two telephone sets.

World Wide Web – two-way, asymmetric data communication over the Internet allowing users' Web browsers to download and present information from Web servers located throughout the world, and allowing users to upload data to the servers.

The general technical requirements and characteristics for these applications are given in Table 4.1. Much of the data was adapted using [3-1].

Table 4.1
Applications -- User Channel Characteristics

End User Application	Min Average User Data Rate per Channel (Mbps)	Max Average User Data Rate per Channel (Mbps)	Latency (sec)	Bit Error Rate (10^{-x})	Connection Type (PTP, Broadcast)	Symmetry (A or S)	Variable/ Continuous
Broadcast Audio	0.256	0.256	30	6	Broadcast	A	C
Broadcast Video	2	6	30	6	Broadcast	A	C
HDTV	8	24	30	6	Broadcast	A	C
Near video-on-demand	2	6	30	6	Broadcast	A	C
Distance Learning	0.128	16	30	6	PTP	A	C
Remote Monitoring	0.064	2	30	6	PTP	A	C
Point of sale transactions	0.064	0.064	1	10	PTP	A	V
World Wide Web	0.064	0.4	0.1	6	PTP	A	V
On-line Internet Service	0.064	0.4	0.1	6	PTP	A	V
LAN-to-LAN Connections	0.386	51.84	0.1	6	PTP	S	V
Audio-on-demand	0.256	0.256	0.3	6	PTP	A	C
Video-on-demand	2	6	0.3	6	PTP	A	C
Video Teleconference	0.128	2	0.3	6	PTP	S	V
Voice Telephony	0.016	0.064	0.1	6	PTP	S	V
Voice Conferencing	0.016	0.064	0.1	6	PTP	S	V
Electronic Gaming	0.064	2	0.1	6	PTP	S	V
Internet File Transfer	0.064	0.386	N/A	6	PTP	A	V
Email	0.064	N/A	N/A	6	PTP	A	V
Information Services	0.064	0.386	N/A	6	PTP	A	C
Database mirroring	2	51.84	N/A	10	PTP	A	V
Software Distribution	0.064	2	N/A	10	PTP	A	V
Audio Burst	0.064	0.128	N/A	10	PTP	A	V
Video Burst	2	51.84	N/A	10	PTP	A	V
File Transfer	0.064	2	N/A	10	PTP	A	V
Medical Imaging	2	51.84	N/A	10	PTP	A	V

2.2 NEXT-GENERATION BROADBAND SATELLITE SYSTEMS

Table 2.2 is a partial compilation of next-generation US Ka band broadband satellite systems. It shows that aggregate satellite data rate capacities fall in the range of 2 to 4 Gbps. Most of these systems orbit at GEO altitudes and are expected to be operational circa Year 2000. Internationally, there are as many as 43 other planned systems.

Table 4.2
Next Generation Ka Band Satellite Systems (FCC Filings)

	Orbit	No. of satellites	User channel data rate, R (Mbps)	No. of channels per satellite, N	Total user data rate, RxN (Gbps)
AT&T VoiceSpan	GEO	12	0.386	8333	3.22
Lockheed Martin Astrolink	GEO	9	0.386	10000	3.86
GE American GE*Star	GEO	9			
Motorola Millenium	GEO	4	0.016	250000	4.00
Hughes Spaceway/Galaxy	GEO	17	0.386	5882	2.27
Morningstar	GEO	4	4	150	0.60
Loral Cyberstar	GEO	3			(2.25 GHz)
Echostar	GEO	2			
KaStar	GEO	2			(6 GHz effective)
NetSat 28	GEO	1	1.544	500000	
Orion F-7,8,9	GEO	3			
PanAmSat 10,11	GEO	2			
VisionStar	GEO	1	4	100	0.40
Motorola Celestri	LEO	63			
Teledesic	LEO	200+	0.016		2.00

Note that AT&T VoiceSpan has been cancelled as of May, 1997.

2.3 GENERATION-AFTER-NEXT SERVICES

Examination of the single-user channel characteristics for the identified applications leads to the observation that there are major logical groupings with common characteristics. The first distinction that we used to group the applications was between broadcast and point-to-point applications. These have very different network architectures. The scaling of the network in terms of the number of channels is also clearly different. The broadcast network scales in relation to the number of programs being broadcast, whereas the point-to-point network scales in accordance with the number of terminal points in the network.

Within the broadcast-type systems, we believe that there are two distinct broadband services, standard (fixed schedule) broadcast television service and video-on-demand service. Both require satellite systems with a large data rate capacity that pushes technology, and both will be major commercial service types in the generation after next.

Examination of the single-user channel characteristics for the identified applications, as well as the service descriptions for the planned next-generation broadband satellite projects, leads to a hypothesis of the following three basic service types for the generation-after-next.. Note that latency is defined differently for the different services. Since there is no interactivity in the broadcast service, the latency is the delay in the network path. For the video-on-demand service, the latency is the delay between a user's request for an action, and the execution of the action. This includes primarily the pause and replay functions which are intended to mimic the user controls on a video cassette recorder. For the bandwidth-on-demand service, latency is the delay between a request for an object to be downloaded and its arrival at the user's terminal.

Broadcast Television Service

Architecture:	One-way broadcast architecture, single or multiple beam.
Quality of Service (QoS):	Near-real-time continuous streaming
User channel data rate:	2-6 Mbps (video component)
Latency:	30 seconds
User bit-error-rate (BER):	10^{-6} typical

Video-on-Demand Service

Architecture:	Single hub to multiple beam content delivery plus user-to-hub uplink/upstream signaling
QoS:	Near-real-time continuous streaming
User channel data rate:	2-6 Mbps (video component)
Latency:	1 second
User BER:	10^{-6} typical

Bandwidth-on-Demand Service

Architecture:	Symmetric, single-hop, user-to-user, or asymmetric hub-to-user downlink
QoS:	Real-time streaming Real-time block transfer Non-real-time
User channel data rate:	16 kbps to 51.84 Mbps and higher
Latency:	0.1 second on real-time block transfer channels 1 second on real-time streaming channels 30 seconds on non-real-time channels
User BER:	10^{-10} typical

3. HIGH LEVEL REQUIREMENTS

This section discusses high level requirements for the hypothetical projected satellite communications services. There are two approaches to deriving requirements, top-down, and bottom-up. The top-down approach uses market research and estimates customer needs to arrive at numbers for sizing (e.g., data capacity) and quality (e.g., latency). The bottom-up approach considers what technology can provide, and fits the application to the technology. Here, we propose satellite requirements based on the top-down approach. We consider how future commercial service providers will seek to compete with equivalent ground-based services. The assumption is that the space-based service must be competitive or better than the ground-based service.

We discuss two top-down requirements for each service that are key drivers of the satellite architecture: the data rate capacity required to support the projected customer base; and, the latency needed to meet the customer's expectations of the system's response.

3.1 BROADCAST TELEVISION SERVICE

3.1.1 Total Average User Data Rate

In order to arrive at a data rate capacity for this service, it is necessary to estimate the desired number of broadcast channels. This is the total number of channels, irrespective of the number of satellites needed to carry these channels. In the broadcast service, there is only a downstream (broadcast source-to-user) signal propagation. All users receive the same signal.

It is difficult to gauge how much the growth in required data rate capacity for future satellite systems will be offset by improvements in compression and frequency reuse techniques. For example, the state of the art in compression of video advances fairly rapidly. On the other hand, compatibility with ground networks and their large installed base may "lock in" certain widely used compression standards, such as MPEG2, for more than one generation.

We will assume that MPEG2 will continue to be the standard in the generation-after-next, and will assume a nominal 4 Mbps MPEG2 channel (average user data rate) to support a broadcast-quality video signal. For simplicity we ignore the audio component of the television signal.

An example of a current generation digital direct broadcast satellite (DBS) system is the Hughes DirecTV system consisting of three satellites, DBS-1,-2, and -3. Each satellite has 16 transponders with a bandwidth of about 30 MHz each. Using a spectral efficiency of 1 bit per second per Hertz, this allows 30 Mbps per transponder. Programming content consumes between 4 Mbps and 6 Mbps per channel, depending on content type, so that each satellite can deliver between 80 and 120 channels. Currently, DBS-1 is running 16 transponders in low-power mode (120 Watts EIRP), while DBS-2 and -3 are running 8 transponders each in high-power mode (240 Watts EIRP). The maximum number of channels for the system is therefore about 256. [3-2]

It is natural to assume that follow-on generations of DBS systems will provide more channels. There are economic arguments for and against this assumption. The main limitation in the television business is availability of programming. The availability of premium quality movies seems to remain relatively constant over time. The same holds true for other types of mass-audience television programming. Increasing the number of channels might therefore require a shift to more specialized or regionalized programming aimed at smaller audiences. For example, DBS providers currently have a competitive disadvantage with respect to cable TV providers in that they do not carry local TV news; but providing these channels means smaller audiences per channel. There is a return on investment that must be calculated to justify these smaller audience channels.

If other types of content, such as block transfers for Web-browsing, are brought into the picture, then it is easy to see how additional capacity can be utilized. The satellite data capacity for the Web-browsing application is sized in proportion to the number of subscribers or users, not in proportion to the number of content-bearing channels. The broadcast service is not optimized for this kind of application. We assume that in the generation-after-next, broadcast satellite architectures will be clearly differentiated from the bandwidth-on-demand architectures, and that future broadcast satellites will not carry Internet service. The use of DBS for Internet service is therefore seen as an interim approach that will eventually disappear.

An innovation requiring increased data capacity is HDTV. HDTV requires 4 times the data rate of NTSC video. It is not clear when HDTV will reach a significant level of market penetration, and we assume that for the generation-after-next satellite systems, HDTV will still constitute only a small percentage of the total channels.

Taking these factors into consideration, we believe that in the generation-after-next a factor 2 increase in the number of channels per system is more likely than an increase by a factor of 5 or 10. The growth of data rate capacity per generation is slower for this type of service than for the bandwidth on demand service because it depends on the number of programming channels, not the number of subscribers. A doubling of capacity would result in a fixed-schedule service consisting of 500 channels, with an average data rate of 4 Mbps each.

3.1.2 Latency

Since there are no real-time response requirements for the television broadcast service, the latency can be relatively large. We will assume the 30 second value for allowable latency, taken from Table 4.1.

3.2 VIDEO-ON-DEMAND SERVICE

3.2.1 Total Average User Data Rate

VOD has been widely held to be a major commercial application in the future. A true VOD service using MPEG2 requires a data rate of about·4 Mbps per user for 2 hours. In recent market-trial systems the channel is dedicated to a single user because the user has full control over the play, pause, reverse, playback, and fast-forward functions. We will define this as true video-on-demand. This requires low latency, because the viewer using his remote control device wants to see nearly instantaneous response. The market trials have shown, however, that this implementation of true VOD is not cost-effective in terms of the number of channels needed, and does not provide enough value-added to make customers be willing to pay for it. [3-3] However, there are other ways to implement VOD-like services, as described next.

The rapid pace of development of solid state memory will eventually make storage of parts or all of a video package in the set-top box cost-effective. This will open up possibilities for alternatives to VOD. One alternative is that the video packages could be burst-downloaded in broadcast mode during non-peak hours. This would take advantage of the fact that viewing is concentrated heavily in certain peak hours (in the evenings on weekends).

Another widely-accepted alternative to VOD is near video-on-demand. A large set-top box storage capacity could allow for an enhanced NVOD service that could substitute for VOD service at a greatly reduced number of content-delivery channels. NVOD today usually means broadcast of programming starting every 5 or 10 minutes. Assume a future enhanced NVOD service broadcasting with start times every minute. A user who selects a package will have the broadcast channel starting on the next 1-minute start time selected by the set top box from the multiplexed bit stream. If he chooses to manipulate the stream using his remote control, the set-top box buffers a video clip of sufficient duration to simulate all the playback functions. When the user resumes regular viewing, the set-top box must select the appropriately timed video stream for display. The number of satellite content-delivery channels for this kind of service is reduced from one channel per user to 120 channels per 2-hour video package; 120 channels at 4 Mbps each totals to 480 Mbps per video package offered. Latency requirements are also mitigated because the real-time functions are performed locally.

A service such as this is a premium service, and might be reserved for the best titles offered by the service provider. A reasonable number for the premium titles offered during any week might be 5, analogous to a movie theater complex showing first-run movies. (The number of available first-run movies is limited in any case.) This assumption results in a total content-delivery data rate of:

Total average user data rate = 5 • 480 Mbps = 2.4 Gbps

Henceforth in this paper, VOD will mean the NVOD implementation of it defined above.

3.2.2 Latency

The low network latency requirement for a pure VOD service does not apply to the hypothetical enhanced NVOD service, if the real-time pause/playback functions are accomplished by buffering the video stream in the set-top unit. With local buffering, the latency is only the latency of the local electronics. Therefore, the round-trip delay to a satellite at GEO altitude, about 0.25 seconds, is acceptable for this service. The content delivery channels would be downlinked to a nation-wide footprint, so a GEO satellite is effective from this perspective as well.

3.3 BANDWIDTH-ON-DEMAND SERVICE

3.3.1 Total Average User Data Rate

Bandwidth-on-demand service will include applications like Internet access and Web browsing for the mass consumer market. The discussion in this section is based on Internet applications as the prototype for the bandwidth-on-demand service. It is difficult to predict what the Internet will look like 10 or more years ahead, but certain points seem clear. In order to compete with traditional video media, the sophistication of the content on the Internet will increase. The component consisting of video and 3-D motion graphics will greatly increase. Video resolution will increase. Some of this increase will be offset by compression, especially the compression of 3-D motion graphics. The 3-D motion component has potentially higher compressibility than ordinary video since the generator application can be transmitted instead of the video output. This is reflected in the MPEG4 standard, which allows for a mix of highly compressed 3-D motion graphics and standard video.

Generation-after-next satellite systems will compete with terrestrial networks to provide Internet services with equivalent data rate capacity and latency.

We will describe a simple equation for Internet service that will allow estimation of the expected traffic in a future satellite system. The input parameters are:

Data rate, r_i , of modem type i
Fraction of users with a given modem data rate, k_i
Intra-session duty cycle (the fraction of time during an on-line session the modem is operating at its data rate), d
Number of subscribers in a service area (a metropolitan service area is assumed), n
Fraction of subscribers on-line in the peak busy period, f

The equation calculates the average data rate of all the modem types in use, and multiplies by the duty cycle to get the average data rate of all on-line sessions. This average rate is multiplied by the number of subscribers times the fraction of subscribers on line, to get the total average user data rate in the service area. The resulting equation is:

Total average user data rate = $(\Sigma(r_i \bullet k_i)) \bullet d \bullet n \bullet f$

Pending further research, values were chosen for the inputs that are essentially informed engineering guesses for the year 2007. These assumptions are:

Average modem data rate for all modem types = $\Sigma(r_i \bullet k_i)$ = 1.38 Mbps, as derived in Table 4.3
Intra-session duty cycle, d = 0.15
Number of subscribers, n = 500,000
Fraction of subscribers on-line, f = 0.1

Table 4.3
Average Modem Data Rate, 2007

Data Rate (bps)	%	Cumulative %	Weighted
9600	1	1	96
14400	1	2	144
28800	2	4	576
56000	5	9	2800
64000	1	10	640
128000	10	20	12800
386000	20	40	77200
1544000	25	65	386000
2000000	25	90	500000
4000000	10	100	400000
Weighted Average (bps)			1380256

The result is 10.35 Gbps total average user data rate in the downstream (Web site-to-user) direction.

3.3.2 Latency

The lowest-latency application carried by a bandwidth-on-demand service is Web browsing, since it involves real-time block-transfers. Web browsing requires low latency because the user must perceive that the time to download an object be almost instantaneous, i.e., less than 100 milliseconds [3-4]. Because of the 0.25 second round-trip delay, GEO-based transport services will therefore be less competitive for this type of application than LEO-based services, since LEO round-trip delays are typically comparable to terrestrial delays. It should be noted however, that network induced delays may dominate over propagation delays in LEO networks or in terrestrial networks. This network induced delay is attributable to many factors in the design and tuning of the network.

GEO satellites are not ruled out for applications not requiring subjectively instantaneous response. This includes real-time streaming applications, and non-real-time applications.

3.4 REQUIREMENTS MATRIX

The requirements derived in the previous sections to support the selected services are summarized in Table 4.4.

Table 4.4
High Level Requirements

	Broadcast Television Service	Video-on-Demand Service	Bandwidth-on-Demand Service
Average User Data Rate (Mbps)	4	4	0.2
Number of User Channels	500	600	50000
Total Average User Data Rate (Gbps)	2	2.4	10
Coverage Area	National	National	Metro
Latency (seconds)	30	1	0.1
User Bit Error Rate	10^{-06}	10^{-10}	10^{-10}

4. SATELLITE PAYLOAD ARCHITECTURE DESIGNS

4.1 INTRODUCTION

This section discusses satellite payload architectures that can be used to support the broadband services described in Section 3. Several architectures are described ranging from those that rely on conventional, state-of-the-art technology, to those that require the introduction of advanced technology and technology that has not yet been space-qualified. State-of-the-art SATCOM technology has been demonstrated through ACTS and other experimental satellites.

Pending a more detailed analysis of traffic and other requirements (e.g., connectivity and transmission quality) for the services, only top-level satellite payload architectures were generated to show the relevant technology drivers. Detailed technical designs based on the satellite payload architectures and detailed tradeoffs among the architectures were not performed. Accordingly, specific parameters required for each technology to support each architecture and each service were not generated.

However, the following sections address issues and assumptions associated with the generation and design of future satellite payload architectures. Design choices and assumptions made include: satellite orbits; operating frequencies; frequency reuse; satellite connectivity; satellite bit-rate capacity versus modulation, satellite beam size, and the receive earth station size to be used.

4.2 TECHNICAL DESIGN ISSUES AND ASSUMPTIONS

4.2.1 Satellite Orbits

Broadband services can be provided using communication satellites located in GEO or non-geostationary earth orbits (NGEOs) such as LEO, MEO and HEO (highly elliptical orbit).

Each GEO or NGEO satellite system has its own advantages and disadvantages. A NGEO satellite system, as compared to a GEO satellite system, requires much less transmit power from its satellites or earth stations, due to its much shorter range to the satellite. It also, in general, provides less transmission delay, which may be critical to certain applications such as interactive voice or interactive data that may be part of the bandwidth-on-demand service. On the other hand, a NGEO satellite system requires many more satellites to be operated in a coordinated fashion for routing signals between its sources and sinks, and may also require tracking antennas. It also restricts the sharing of the natural frequency spectrum with other systems (GEO or NGEO), due to radio frequency (RF) interference [4-1, 4-2].

This report is limited to describing satellite payload architecture designs using GEO satellites. Designs using NGEO satellites are left for future studies.

4.2.2 Satellite Operating Frequencies

To make full use of the Natural Frequency Spectrum, the International Telecommunications Union (ITU), an organization of the United Nations:

- Has classified radio emissions into services such as BSS (Broadcast Satellite Services), FSS (Fixed Satellite Service), MSS (Mobile Satellite Service), BS (Broadcast Service), and FS (Fixed Service).
- Allocates frequency bands to these services.
- Sets priorities for these services on a specific frequency band (i.e., whether the services are primary or secondary), and sets constraints and coordination rules on these services to prevent them from causing harmful interference to each other [4-3].

The relevant ITU services for the three broadband services considered in this report are: BSS, FSS and MSS. BSS [4-3] is a radio communication service in which signals transmitted or retransmitted by satellites are intended for direct reception by the general public. FSS [4-3] is a radio communication service between earth stations at specified fixed points when one or more satellites are used. And MSS [4-3] is a radio communication service between mobile earth stations and one or more satellites, or between satellites used by this service or between mobile earth stations by means of one or more satellites.

The TV-broadcast (TVB) service is a BSS; the video-on-demand (VOD) service, depending on implementation, may be a BSS or an FSS; and the bandwidth-on-demand (BWOD) service, depending on usage of the bandwidth, may be an FSS or an MSS. In general, the ITU constraints are much more relaxed for a BSS or an MSS than for an FSS to allow the use of much lower cost, smaller sized earth stations. For instance, in the US, Ku-band BSS GEO satellites are spaced at 9° apart (instead of 2° apart as set for Ku-band FSS GEO satellites) and the constraints on satellite effective isotropic radiated power (EIRP) and on satellite EIRP density for Ku-band BSS GEO satellites are much more relaxed than those of Ku-band FSS satellites.

Table 4.5 shows the commercial (non-government) frequency allocation in the US [4-4]. Table 4.6 provides a summary of bandwidth available for the three ITU services (BSS, FSS and MSS) that can be used to support the three broadband services (TVB, VOD and BWOD). It is noted that:

- The FSS Ku-band in the US, with uplink at 14.0 - 14.5 GHz and downlink at 11.7 - 12.2 GHz, is dedicated to satellite services (i.e., there is no terrestrial interference). However, this frequency band is heavily used by many different GEO satellites serving the US (e.g., GE's Gstar, Spacenet, GE, and SATCOM; ATT Skynet's Telstar; and Hughes' SBS, and Galaxy).

Table 4.5
Commercial (Non-Government) Frequency Allocation in the USA.

Frequency Band, GHz	Uplink (Earth-to-Space)	Downlink (Space-to-Earth)
Ku-Band		
10.7 - 11.7		FSS , FS, MS, SRS
12.2 - 12.2		FSS
12.2 - 12.7		BSS
12.70 - 12.75		FSS , FS, MS
12.75 - 13.25	FSS, FS, MS	
14.0 - 14.2	FSS, RNS	
14.2 - 14.5	FSS	
17.3 - 17.7	FSS	
17.7 - 17.8	FSS, FS, MS	
Ka-Band		
17.7 - 17.8		FSS, FS, MS
17.8 - 18.6		FSS, FS, MS
18.6 - 18.8		FSS, EESS, FS, MS, SRS
18.8 - 19.7		FSS, FS, MS
19.7 - 20.2		FSS, MSS
27.5 - 29.5	FSS, FS, MS	
29.5 - 30.0	FSS	
50/40 GHz Band		
37.5 - 38.0		FSS, FS, MS, SRS
38.0 - 39.5		FSS, FS, MS
39.5 - 40.5		FSS, MSS , FS, MS
40.5 - 42.5		BSS
42.5 - 43.5	FSS, FS, MS, RAS	
45.5 - 47.0	MSS, MS, RNS, RNSS	
47.2 - 50.2	FSS, FS, MS	
50.4 - 51.4	FSS, FS, MS	

Definitions

FSS: Fixed Satellite Service
BSS: Broadcast Satellite Service
MSS: Mobile Satellite Service
RNSS: Radio Navigation Satellite Service
EESS: Earth Exploration Satellite Services

RAS: Radio Astronomy Service
FS: Fixed Service (non-satellite)
MS: Mobile Service (non-satellite)
RNS: Radio Navigation Service
SRS: Space Research Service

Table 4.6
Commercial (Non-Government) Bandwidth Available for BSS, FSS, and MSS at Ku-Band, Ka-Band, and 50/40-GHz Band in the USA

	Ku-Band	Ka-Band	50/40 GHz Band
BSS	500 MHz	Not Allocated *	2 GHz
FSS	500 MHz	2.5 GHz (max.)**	3 GHz (max.)***
MSS	Not Allocated	500 MHz (max.)	1 GHz (max.)

* 21.4 - 22.0 GHz was allocated to BSS in Region 1 and Region 3 (US in Region 2).

** 27.5 - 29.5 and 29.5 - 30.0 GHz (2.5 GHz of bandwidth) can be used on the uplink, but 17.7 - 20.2 GHz (3 GHz) is available for the downlink FSS. (from Table 4.5).

*** 5 GHz of bandwidth is available on the uplink, but only 3 GHz of downlink bandwidth is available (see Table 4.5).

- The BSS Ku-band in the US, 12.2 - 12.7 GHz, is allocated with the corresponding feederlink FSS Ku-band 17.3 - 17.8 GHz.
- There are no Ku-band allocations for the MSS.
- There are no Ka-band allocations for the BSS.
- There are no Ka-band allocations for the MSS uplink.
- There are currently no commercial satellites in the US operated at Ka-band or the 50/40 GHz band. However, there have been at least 13 FCC filings for satellite systems to be operated at FSS Ka-band (e.g., Celestri, and Teledesic) [4-5] and 2 FCC filings for satellite systems to be operated at FSS 50/40 GHz band (M-Star [4-6] and V-Stream [4-7]).

Because the Ku-band is quite congested, it is likely that the proposed broadband services will be operated at Ka-band or 50/40 GHz. Right now, the Ka-band and 50/40 GHz band technology is still rather new and is being evaluated on an experimental basis in the NASA ACTS program, the Italian ITALSAT program, and the Japanese Engineering Technology Satellite (ETS) program. However, the technology will become more mature after the launch and operation of the next generation FSS Ka-band and 50/40 GHz band satellites. Another possibility is to use the optical frequency band [4-8], [4-9], if the technology is feasible and cost effective by the time the broadband services are launched. Key issues with respect to optical bands are cloud cover limitations over many areas of the US and the resultant impacts on link attenuation and signal scattering losses (See Section 5 for a brief discussion of laser technology for earth-space communication.)

Use of Ku-band and particularly of Ka-band and 50/40 GHz band are greatly affected by rainfall at the transmit and receive earth station locations. Rainfall attenuates and depolarizes radio frequency signals. It also significantly raises the noise temperature of the receive earth station [4-10], [4-11]. Table 4.7 shows the rain attenuation characteristics, link availability, expressed in percent of a typical year that link fade level will be less than the indicated amount. The results of Table 4.7 have been generated from the Crane model [4-11], at 12 GHz (representative Ku-band downlink frequency), 20 GHz (representative Ka-band downlink frequency) and 40 GHz (representative 50/40 GHzband downlink frequency) for three different representative rain regions: Crane rain zone F (dry region, e.g., Los Angeles); Crane rain zone D2 (medium region, e.g., New York); and, Crane rain zone E (wet region, e.g., Miami). The results for the first three columns of Table 4.7 are based on typical rain zones assuming a nominal latitude of 40° North, and a nominal terminal to satellite elevation angle of 20°. The use of a common set of parameters help to display the impacts of different rain rate regions, however, in actuality, the southern Florida area (Miami) is at approximately 30°N, and this area would not typically have a 20° elevation angle from a US-only coverage GEO satellite (see for example the figure below). Therefore the fourth column of Table 4.7 has been added to show a more realistic link attenuation model for southern Florida (Miami) area, which would, for a satellite typically centered over the US, have a nominal terminal to satellite elevation of approximately 45°.

From the first three columns of Table 4.7 (with a constant set of parameters except for rain zone type), the downlink rain attenuation in decibels (dB) increases about 3 times from Ku-band to Ka-band and about 10 times going from Ku-band to 50/40 GHz band (i.e, on the order of the frequency ratio to a power, $(f_{upper}/f_{lower})^{2.2}$. Thus, due to the very large attenuations that can be experienced at Ku and particularly, 50/40 GHz, it is necessary to employ techniques that can mitigate such losses or actively compensate for rain attenuation, i.e., adding link power from a reserve pool of power when required for a particular link

Table 4.7
Rain Attenuation (dB) Versus Crane Rain Zone, Frequency, Availability and Outage Period

Availability (%)	Outage per year (hours)	Zone F (LA)*			Zone D2 (NY)*			Zone E (Miami equivalent)*			Zone E (Miami)**		
		12	20	40 (GHz)	12	20	40 (GHz)	12	20	40 (GHz)	12	20	40 (GHz)
95.0	437	0.00	0.00	0.00	0.00	0.00	0.00	0.15	0.51	2.49	0.10	0.32	1.64
96.0	349	0.02	0.07	0.40	0.12	0.40	2.04	0.32	1.05	4.79	0.22	0.71	3.32
97.0	262	0.04	0.15	0.85	0.27	0.88	4.12	0.55	1.75	7.57	0.38	1.22	5.36
98.0	175	0.08	0.28	1.47	0.49	1.57	6.90	0.88	2.76	11.40	0.61	1.94	8.11
99.0	87	0.28	0.92	4.31	0.99	3.11	12.82	1.81	5.58	21.53	1.30	4.0	15.47
99.5	44	0.56	1.80	7.96	1.73	5.35	20.89	3.15	9.66	35.30	2.31	6.99	25.47
99.8	17.5	1.23	3.85	15.8	3.15	9.58	35.4	6.17	18.32	62.9	4.57	13.5	45.9
99.9	8.7	2.02	6.24	24.4	4.75	14.26	50.8	9.95	29.10	95.4	7.53	21.9	70.6

* Note: First three columns .are for rain zones at 20° Elevation Angle and 40° N Latitude, using Crane Global Model & Marshall-Palmer Rain Drop Size Distribution Model. (The indicated cities are cited for typical climates only.)

** The fourth column provides results for rain zone E near Miami given a satellite location at 85° West, a nominal Miami area latitude of 30°N, and a nominal terminal to satellite elevation of 45°.

rather than statically reserving large amounts of spacecraft power for all links for the few hours of the year in which a very deep fade can occur on a particular link. Examples of these techniques are [4-11]: on-board regeneration, uplink power control, downlink power control, and variable forward error correction (FEC) coding rate. Techniques that utilize the available satellite power efficiently are also necessary. Examples of these techniques are power-efficient modulation/coding techniques (e.g., QPSK with Turbo codes [4-12], [4-13]), and automatic power sharing techniques among the downlink beams (e.g., active transmit phased array (ATPA) antennas [4-14] and the matrix amplifier and routing system (MARS) [4-15], [4-16]). These power-efficient techniques and efficient power-sharing techniques will be discussed later in this report.

Note that Figure 4.1 shows the minimum terminal to satellite elevation for a CONUS centered GEO satellite would typically be grater that 20°. The figure also provides the reader with a sense of the coverage area subtended by satellite beam sizes of 5 an 10 degrees, i.e., nominal coverage areas of 3100 km and 6200 km at the sub-satellite points, which are somewhat larger if focused on the US from 85 W as indicated. Also, for the purpose of discussion later in this report, .a 7.5 degree beam is assumed to be the typical coverage area for CONUS.

The satellite payload architectures to be generated in this section are top-level architectures which can be used with any frequencies: Ku, Ka or 50/40 GHz bands.

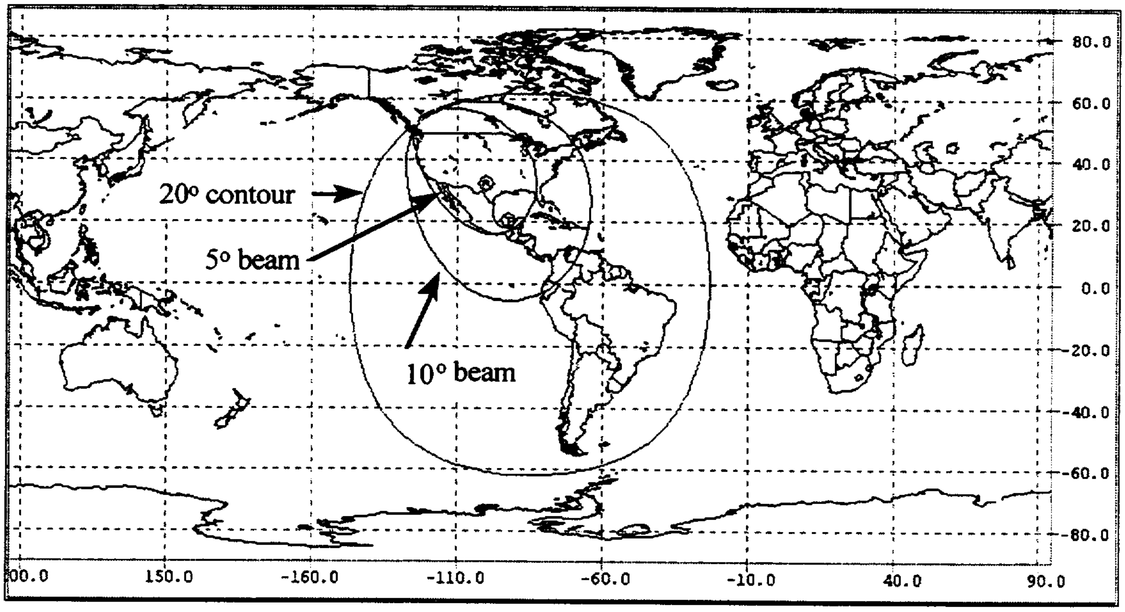


Figure 4.1 Typical US Satellite GEO Coverage

4.2.3 Frequency Reuse

To provide adequate bandwidth for broadband services, it may be necessary to reuse the frequency bands allocated by the ITU and the FCC. Frequencies may be reused by utilizing additional satellites operating at the same frequency band and relying on the earth station antenna's off-axis gain discrimination.

Frequencies may also be reused on the same satellite through spatial diversity of the satellite antenna beams and also through polarization discrimination. The satellite architectures developed in this section of the report do not rely on frequency re-use techniques to simplify the discussion. However techniques and their application to services discussed in this report would provide significant expansion of current capabilities and they are suggested as the subjects for future studies.

4.2.4 Satellite Connectivity

It is necessary to use more than one GEO satellite to provide services to a geographical area wider than a satellite field of view, or to provide services to the same area that requires more bandwidth and power than a single satellite can provide. In order to have full connectivity between the sources and the sinks, it is necessary to connect the GEO satellites together via either intersatellite links [4-17] or gateway earth stations. Use of intersatellite links in general reduces satellite transmission delay but makes the satellite payload architectures more massive and more complicated. This is also suggested for future study.

4.2.5 Satellite Data Rate Capacity

A satellite is designed and operated with its available bandwidth and power. Its bit-rate capacity based on the available bandwidth depends solely on the modulation/coding schemes used. To have highest bit-rate capacity based on the available bandwidth, it is desirable to use bandwidth-efficient modulation schemes, e.g., octal or 8-phase shift keying (8-PSK) or higher order m-ary quadrature amplitude modulation (mQAM) sub-systems. Nevertheless, use of these schemes may require significant additional satellite power. Accordingly, a modulation that uses more bandwidth but less power, e.g., quaternary phase shift keying (QPSK), may be used. Currently, forward error correction (FEC) is typically used to further reduce satellite power required at the expense of satellite bandwidth. For digital video broadcast, modulation/coding schemes have been standardized in Europe using the digital video broadcast - satellite (DVB-S) standard [4-18]. For the DVB-S standard, the modulation is QPSK and the FEC coding is the convolutional code of different code rate concatenated with a Reed-Solomon block code of (188, 204) code rate. Table 4.8 shows the bit-rate capacity of a satellite based on the satellite available bandwidth and the convolutional code rate used, using the standard assumption of 1.4 times the symbol rate for the required satellite bandwidth.

Table 4.8
Supportable Data Rate Vs. Available Satellite Bandwidth

Satellite Bandwidth	DVB-S Standard: QPSK with (188, 204)-RS Code Concatenated with Convolutional Code with Rate of:		
	7/8	3/4	1/2
500 MHz	576 Mbps	494 Mbps	329 Mbps
1 GHz	1.15 Gbps	988 Mbps	658 Mbps
2 GHz	2.30 Gbps	1.98 Gbps	1.32 Gbps
5 GHz	5.76 Gbps	4.94 Gbps	3.29 Gbps
10 GHz*	11.52 Gbps	9.88 Gbps	6.58 Gbps

* Would require significant frequency reuse.

To estimate the satellite bit-rate capacity based on the available satellite power is much more involved, as it depends on many factors including modulation/coding schemes, sizes of satellite and earth station antennas, whether the satellite utilizes onboard regeneration, the quality of information (e.g., BER or Eb/No and link availability) required, and the interference environment the satellite and the earth stations are in.

Table 4.9 summarizes the bit-rate capacity based on the available satellite power for different satellite downlink beam sizes, different receive earth station antenna sizes, and the three frequency bands. The following assumptions were used to generate Table 4.9:

- The total DC prime power available from a satellite bus is 8000 W. This assumption is somewhat high because the HS-601 bus available from Hughes Space and Communications typically provides only about 3000 W [4-19][1]. Nevertheless this assumption is also feasible based on project Omega [4-20], so that the satellite transmitter available power is 1460 W (i.e., 31.6 dBW), based on the following calculations and assumptions:

8000 W	Total DC prime power (project Omega)
670 W	Power required for onboard processing (from ACTS [4-21])
30 W	Power required for other components (LNA, downconverters, etc.)
7300 W	Power available for satellite HPAs
2920 W	HPA RF power (assuming 40 % efficiency)
1460 W	Power available from satellite transmitter (31.6 dBW; assuming 3 dB for output backoff and output circuit loss).

- The satellite utilizes onboard regeneration with the DVB-S standard: QPSK with a 1/2-rate convolutional code concatenated with a (188,204) Reed-Solomon code.

Table 4.9
Supportable Data Rate Vs. Downlink Beam Size and Rx Earth Station Size for Ku-Band, Ka-Band and 50/40 GHz Band*

	Ku-Band		Ka-Band		50/40 GHz Band	
Rx ES Antenna	**0.25° Beam (100 Mile Spot)**	**7.5° Beam (Nominal US)**	**0.25° Beam (100 Mile Spot)**	**7.5° Beam (Nominal US)**	**0.25° Beam (100 Mile Spot)**	**7.5° Beam (Nominal US)**
0.15 m	*9.4 Gbps*	10 Mbps	651 Mbps	724 kbps	146 Mbps	162 kbps
0.30 m	*37.7 Gbps*	42 Mbps	2.6 Gbps	3 Mbps	583 Mbps	648 kbps
0.60 m	*151 Gbps*	167 Mbps	*10.4 Gbps*	12 Mbps	2.3 Gbps	3 Mbps
1.2 m	603 Gbps	670 Mbps	*41.7 Gbps*	46 Mbps	*9.3 Gbps*	10 Mbps
2.4 m	*2410 Gbps*	2.7 Gbps	*166.7 Gbps*	185 Mbps	*37.3 Gbps*	41 Mbps

* Notes: values in italics are band limited and would likely be used with broader coverages (i.e., larger beam widths or smaller satellite HPA's). Also, above assumes almost all of satellite resources allocated to one beam area; should divide by number of beams required to support total coverage area. The initial conclusion is that small terminals, e.g., 0.15m (6 inch) up to 0.6m (2-ft), require small beam coverages, approaching 0.25 degree, to achieve high data rates.

[1] The Hughes HS-701 is projected to provide up to 10 kW depending upon the payload weight requirement.

- The links utilize both uplink and downlink power control with the minimum acceptable downlink Eb/No assumed to be 6 dB.
- The satellite parameters (e.g., antenna gains) and the earth station parameters (e.g., G/Ts) are based on those used in the ACTS experiments [4-21, 4-22] and the ITALSAT experiments [4-23].

A sample downlink budget calculation is tabulated in Table 4.10.

Note that for a GEO satellite, a 0.1° coverage corresponds to 38.8 miles so a 0.25° beam covers roughly an area of 100 miles in diameter; and, a 7.5° satellite beam covers roughly an area of 3000 miles in diameter. (See also Figure 4.1 which shows coverage areas of 5° and 10° for reference.) Also note that there are ITU power density power constraints and adjacent satellite interference problems that may preclude the use of very small earth station antennas (e.g., 0.15 m), particularly at Ku-band. In addition, there are constraints on satellite weight and space that may also preclude the use of many narrow 0.25° beams. Such trade-off issues are left for future studies.

Table 4.10
Sample Downlink Power Budget Calculations*

	Ku-Band	Ka-Band	50/40 GHz Band
HPA RF Power, dBW	31.6	31.6	31.6
Ant. Gain (0.25°), dBi	53.0	53.0	53.0
Path Loss, dB	205.6	211.0	217.0
ES G/T (1.2 m), dB/K	19.2	16.0	21.7
Boltzmann's Constant	-228.6	-228.6	-228.6
D/L Control, dB	3.0	6.0	12.0
C/No, dB-Hz	123.8	112.2	105.7
Required Eb/No, dB	6.0	6.0	6.0
Data Rate, dB-bps	117.8 *(603 Gbps)*	106.2 *(41.7 Gbps)*	99.7 (9.3 Gbps)

* Note: values in italics are obviously band limited and would likely be used with broader coverages (i.e., larger beam widths or smaller satellite HPA's).

4.3 SATELLITE TV-BROADCAST SERVICE

4.3.1 Satellite Network Architecture

Figure 4.2 is a satellite network architecture for the TV-broadcast (TVB) service described in Section 3. In this architecture, encrypted digital TV programs are fed from a broadcast center to the satellite for broadcast to a wide area on earth. In this wide area, subscribers can view their selected TV programs with proper equipment (i.e., off-the-shelf TV receive-only terminals) that are used to receive, decrypt, and display the TV programs.

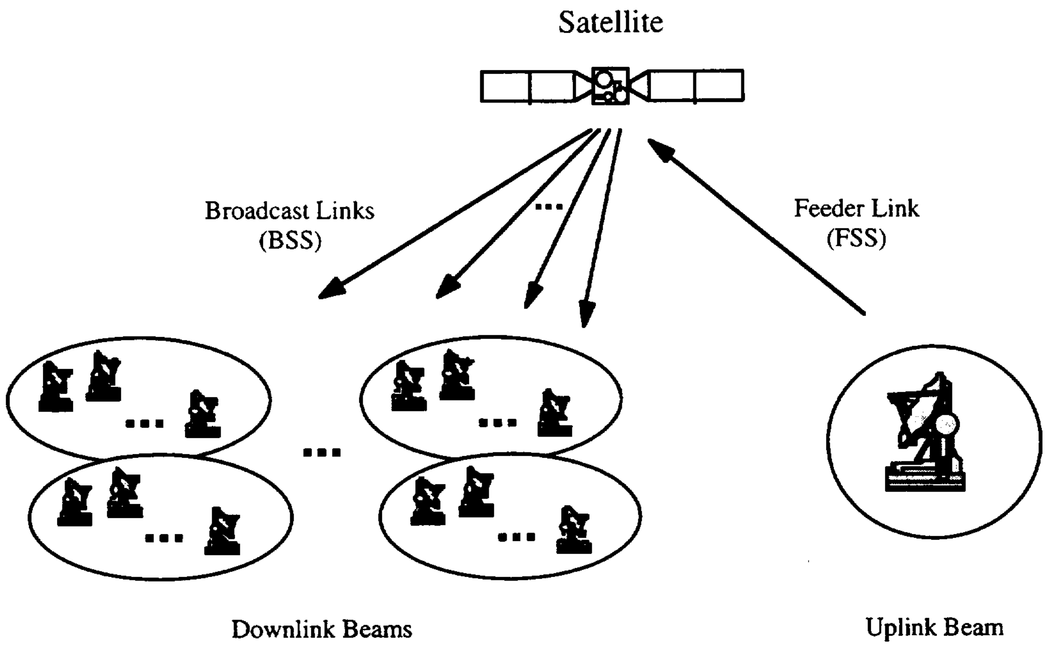


Figure 4.2 Satellite TV Broadcast Service: Network Architecture TVB

This service is qualified to use the BSS frequency bands allocated for satellite broadcast. As discussed earlier in Section 4.2, the advantage of using the BSS frequency bands, over the FSS frequency bands, is that the receive antenna can be much smaller (e.g., 0.45 m at BSS Ku-Band with its ~4° beamwidth, and smaller antennas at Ka-Band or higher frequency bands). Use of small antenna at Ku-band BSS frequencies is possible because BSS Ku-band satellites are separated by 9 degrees, thus avoiding harmful adjacent satellite interference. This may be contrasted with FSS Ku-band GEO satellites which only have 2 degree separations, since they assume the use of larger receive terminals (i.e., terminals with smaller beamwidths). The use of smaller terminals for BSS is also permitted because the satellite EIRP and EIRP density are allowed by the ITU to have much higher values to compensate for lower gains of the receive antennas. One question that remains is whether the use of higher frequency bands such as Ka-band will permit terminals smaller than 4° beamwidths or whether the 0.45 m terminal sizes would be retained to permit smaller BSS Ka-band satellite spacings (i.e., < 9 degrees for Ka BSS satellites). This is an open issue and could result in a compromise to achieve larger system capacity with moderate user terminal size reduction.

From Figure 4.2, there is one uplink (feeder link) beam operated at FSS frequencies and multiple downlink (broadcast link) beams operated at BSS frequencies. In order to reduce DC prime power required from the satellite, the design utilizes multiple narrow beams instead of one wide beam for the downlink to cover the entire area of interest. To optimize usage of the bandwidth available, the same TV programs are broadcast to the entire wide area using the same frequencies; therefore the multiple downlink beams must be cophased to ensure that they do not combine destructively at their beam edges. Nevertheless, to accommodate programs which are sensitive to time (e.g., news), the same program can be updated and broadcast again.

The TV programs can be broadcast using single channel[2] per carrier (SCPC) carriers or multiple channels per carrier/time division multiplex (MCPC/TDM) carriers[3]. The design of the receive terminals with SCPC carriers is much simpler because it allows a single TV program to be transported by a single carrier as implemented using current AM/FM radio broadcast or TV network broadcast carriers. Another advantage of the SCPC carrier design is that each individual TV program can be uplinked from different locations within the uplink beam.

With MCPC/TDM carriers, multiple TV programs are brought in at the same location to be multiplexed together in a time division manner and to be transported by a single carrier. One advantage of using MCPC/TDM carriers is that the data compression of the TV programs can be flexibly shared and compensated for by each other in real time to improve video quality without using additional bandwidth. The total aggregate transmission rate is fixed. MCPC/TDM carriers are currently used by the US direct-TV broadcast-to-home service providers (e.g., DirecTV, USSB and PrimeStar). Formats for compression, multiplexing and modulation, including FEC coding, for DVB-S have also been standardized [4-18].

To optimize satellite power and bandwidth utilization, 8-PSK or 16-QAM modulation together with a Turbo code can be used. As compared to QPSK or BPSK, the two most used digital modulations in satellite communications, 8-PSK uses 2/3 of the bandwidth required by QPSK and only 1/3 of the bandwidth required by BPSK. 8-PSK was developed and field-tested by COMSAT Labs. in 1990 [4-24] to demonstrate that an information rate of 140 Mbps can be transported through an Intelsat Ku-Band transponder with a bandwidth of just 72 MHz. This was done by using 8-PSK with 7/9-rate FEC. 16-QAM improves the bandwidth utilization further by a factor of 3/4 and has been used extensively in terrestrial communications, but has not been used in satellite communications due to uncertainty in path losses and in transponder nonlinearity. A DoD/DISA initiative investigated an alternative DSCS offload approach that would ease the transition of wideband users to other media while at the same time providing substantial support to tactical users on DSCS X-band satellites. For the last two years, a simulation has been performed and field tests have been conducted over a DSCS satellite using 8-PSK and 16-QAM modems, with different FEC, built by various modem manufacturers (e.g., EF Data and Radyne) [4-25]. Technology is now available to test higher order modulation formats that can result in bandwidth compression to levels shown in Table 4.11. Note that 32-QAM has been tested over DSCS X-band satellites. Traditional use of rate 1/2 QPSK with 256-QAM using a rate 3/4 concatenated with a block code of high rate (e.g, rate 0.9) can potentially result in a bandwidth compression of approximately 5.4 over current systems (1.4 Hz/bps / 0.175/(0.75*0.9)).

Turbo codes are a new FEC technology which also offers significant improvement over common conventional convolutional FEC techniques [4-12, 4-13]. Turbo codes were introduced in 1993 with a claim that their performances are very close to the Shannon limit. Since then the claim has been verified and simulation studies, including those at NASA JPL, have been conducted to extend the basic ideas of the Turbo codes for transmission over environments other than the idealistic AWGN environment.

[2] Note that a "channel" as defined from a SATCOM point of view is not a user service channel, but is instead all of the information placed on a transmitted signal; i.e., the "channel" in SCPC terminology includes all user services on each transmitted radio frequency carrier.

[3] For MCPC, the transmitted carrier consists of time slices that can contain distinct user services.

Table 4.11 Bandwidth Requirements by Modulation Type and Order

Modulation type & Order (M)	Required Bandwidth / code rate (r_c)
4 (QPSK)	$0.7 / r_c$
8 (8-PSK)	$0.467 / r_c$
16-QAM	$0.35 / r_c$
32-QAM	$0.28 / r_c$
64-QAM	$0.233 / r_c$
128-QAM	$0.2 / r_c$
256-QAM	$0.175 / r_c$

Although not demonstrated in this report, a significant potential exists for adapting modulation type (e.g., bandwidth compression via use of adaptive modulation formats) in combination with adaptive rate forward error correction (e.g., code rate and type and interleaver depth) to adapt to changing rain attenuation and other transient link degradation effects. This adaptation could be performed in combination with the use of non-constant bit rate channels through the emerging Asynchronous Transfer Mode (ATM) standards. This should be the subject of future studies and is only mentioned here to indicate the potential for effective system operations with higher frequency band technologies that will have to perform in the very large and highly variable link attenuation environments that would be experienced with the future Ka-band systems discussed in the following sections.

4.3.2 Satellite Payload Architectures

To support the generation-after-next TV broadcast service, three satellite payload architectures are addressed here to show the relevant technologies required: Architectures TVB1, TVB2 and TVB3.

Architecture TVB1

The first architecture, Architecture TVB1 shown in Figure 4.3, is a conventional "bent-pipe" architecture where no onboard processing is employed. The satellite only performs frequency translation through a down-converter (D/C), amplifies, and reradiates the carriers back to earth via the downlink beams.

Rainfall severely attenuates and depolarizes carriers at the frequencies considered here for the service (Ka-band or 50/40 GHz band; see Section 4.2). To improve satellite link availability, real-time uplink power control at the transmit earth station is used to compensate for uplink rain fade and more power (EIRP) is allocated to the downlink beams to compensate for the downlink rain fade. Because each downlink beam covers different geographical areas with different rain fade characteristics, the EIRP allocated for each downlink is different even though each beam carries the same TV programs.

Since there is no need to reshape or to steer the beams, and the amount of RF power required to support each beam is fixed, conventional reflector antennas instead of active phase array antennas are used.

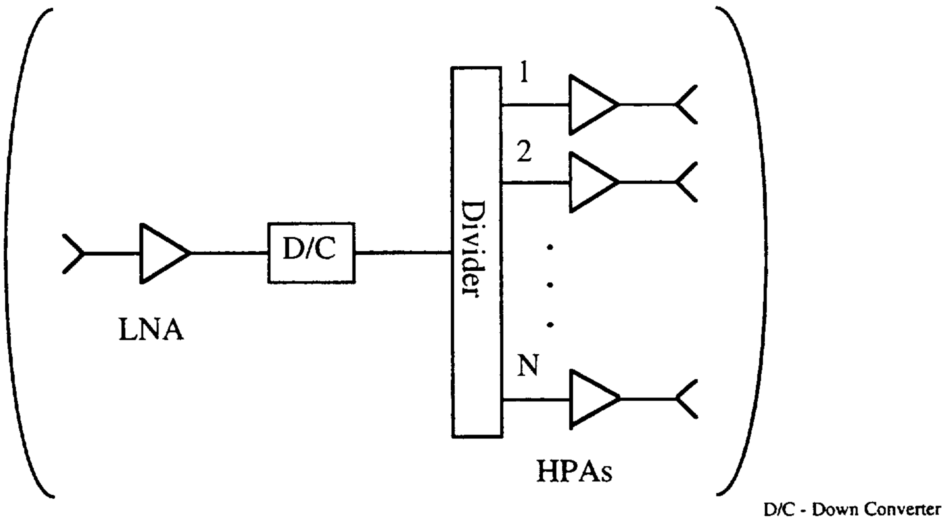


Figure 4.3 Satellite TV Broadcast Service: Satellite Payload Architecture TVB1

Architecture TVB2

The second satellite payload architecture, Architecture TVB2 shown in Figure 4.4, is an onboard regeneration architecture. In Architecture TVB2, the carriers are demodulated and remodulated to decouple the uplink and downlink communication impairment. As a result, the downlink EIRP required is reduced by a few dBs or more, depending on the operating BER requirement, the waveform used, and the payload design implemented. From Figure 4.4, the carriers are down-converted, frequency-demultiplexed (i.e., filtered), demodulated and FEC decoded, remodulated and FEC recoded, remultiplexed, upconverted, amplified, and reradiated to the downlink beams.

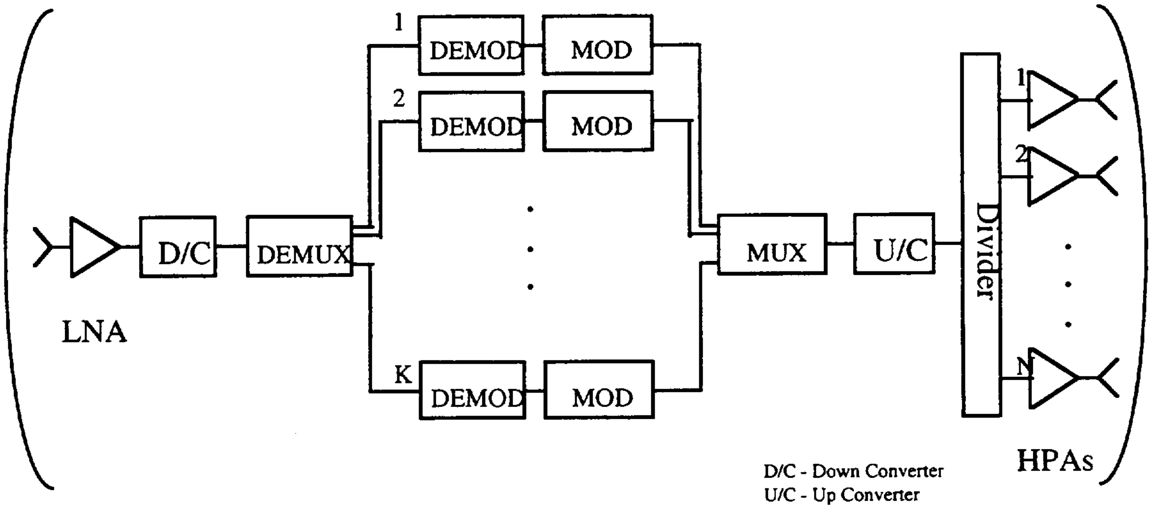


Figure 4.4 Satellite TV Broadcast Service: Satellite Payload Architecture TVB2

The technology of onboard regeneration (i.e., demodulation/remodulation) has been implemented on the ACTS satellite [4-26] and other satellites (e.g., the experimental Italsat, military Milstar).

Onboard multicarrier demultiplexing (FDM) can be carried out with conventional microwave filters or surface acoustic wave (SAW) filters. When the number of carriers is large or when the carrier center frequencies or the carrier bandwidth needs to be adjustable, onboard multicarrier demultiplexing can be performed with a digital processing technology (e.g., FFT/IFFT or polyphase), which has been the subject of research at several institutions, some of which was sponsored by NASA [4-27].

Architecture TVB3

The last satellite payload architecture, Architecture TVB3 shown in Figure 4.5, is the extension of Architecture TVB2. In Architecture TVB3, to further reduce the DC prime power required from the satellite, downlink real-time power control, instead of allocation of downlink power margin, is used to compensate for downlink rain fade. In order to optimally utilize the HPA power available onboard the satellite, Architecture TVB3 utilizes the matrix amplifier and routing system (MARS) [4-15, 4-16]. With the MARS technology, the power outputs from all HPAs are automatically pooled together and automatically shared among the downlink beams, regardless of the amount of power used by each beam. With the MARS technology, the power of signals from each MARS input port can be routed or distributed equally or unequally to one or more of the output ports (e.g., downlink beams).

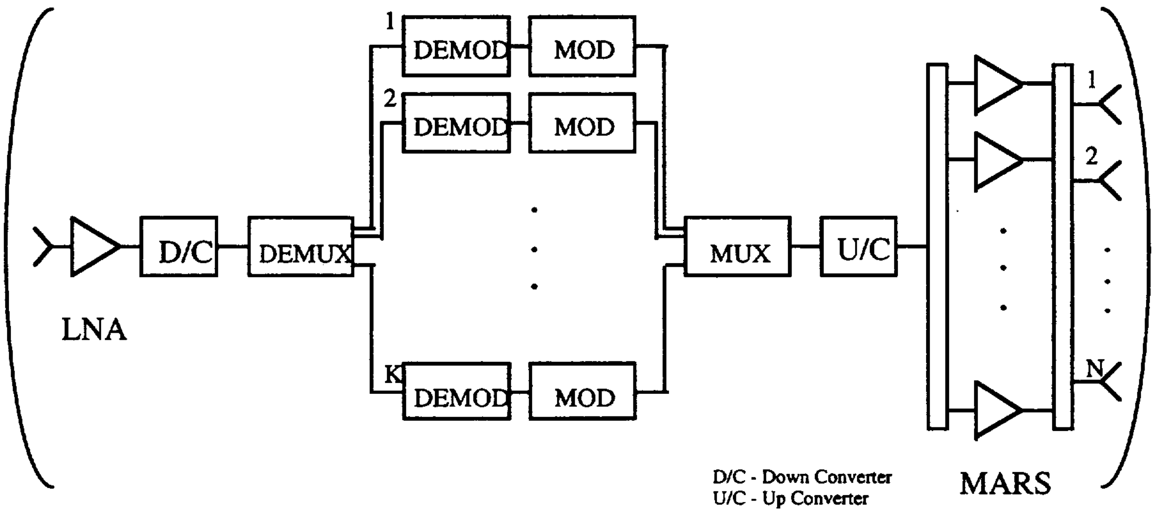


Figure 4.5 Satellite TV Broadcast Service: Satellite Payload Architecture TVB3

The MARS technology was developed at SAIC in 1993 under the sponsorship of the US Air Force [4-15, 4-16] in response to the need for increased power efficiency in communication satellites. It is a generalization of the matrix amplifier, an RF power sharing technology which was first conceived at COMSAT Laboratories in 1972 as a Butler Matrix transponder [4-28]. The MARS technology also extends the power sharing concept by adding the routing capability to one or more output ports without using any multiplexers or switches placed after the HPAs. Although many general characteristics (e.g., the construction and realization of a MARS by different microwave components, MARS nonlinearity/intermodulation effects, MARS routing and distribution) have been identified [4-15], some characteristics are not fully understood and need further investigation, particularly in the area of signal routing and power distribution.

Note also that there may be an issue of cophasing the beams when MARS is used. This is because the phase of each carrier coming out of each downlink beam varies with respect to the setting of the MARS power distribution to the downlink beams. This issue needs to be investigated if Architecture TVB3 is utilized.

4.3.3 Supporting Technology Summary

The supporting technologies required for the three satellite payload architectures for the satellite TV broadcast service are summarized in Table 4.12.

Table 4.12
Supporting Technology for the Satellite TV Broadcast Service

Satellite Payload Architecture	Supporting Technology	Comments
TVB1	• Largely Conventional • (See Below)	• Baseline for comparison
TVB2	• Onboard Regeneration • Onboard Demultiplexer	• Decouples uplink and downlink impairments • Demultiplexer and multiplexer (MUX) can be implemented using conventional microwave filters or Surface Acoustic Wave (SAW) technology • Reduces satellite and terminal transmit RF power requirements and, hence, DC prime power requirements
TVB3	• Onboard Regeneration • Onboard Demultiplexer • MARS	• Same as for TVB2 , but permits greater adaptivity with the larger fade dynamic range associate with higher frequency band (Ka-band & 50/40 GHz) operations
Common (TVB1, TVB2, TVB3)	• Ka-band or 50/40 GHz Technology • Video Compression • Combined Video Compression/Video TDM • M-PSK and M-QAM Modulation with Turbo Codes	• All options can take advantage of higher frequency bands • Can utilize greater adaptivity of modulation formats and FEC to counter increased fade depth experienced at higher frequency bands (e.g., at Ka-band and 50/40 GHz)

4.4 SATELLITE VIDEO-ON-DEMAND SERVICE

4.4.1 Satellite Network Architecture

There are two types of communication links required for the satellite video-on-demand (VOD) service described in Section 3: delivery links and signaling links.

The delivery links are one-way links that are used to deliver video signals to appropriate subscribers upon request. The architecture for the delivery links, Architecture VOD-D shown in Figure 4.6, has multiple downlink beams and one uplink beam as in Architecture TVB in Section 4.3. The only difference is that each downlink beam of Architecture VOD-D receives different signals and therefore there is no need to cophase the beams. From the spectrum allocation standpoint, it is not clear whether the service still qualifies as a BSS service. More likely the service will be implemented using FSS frequencies.

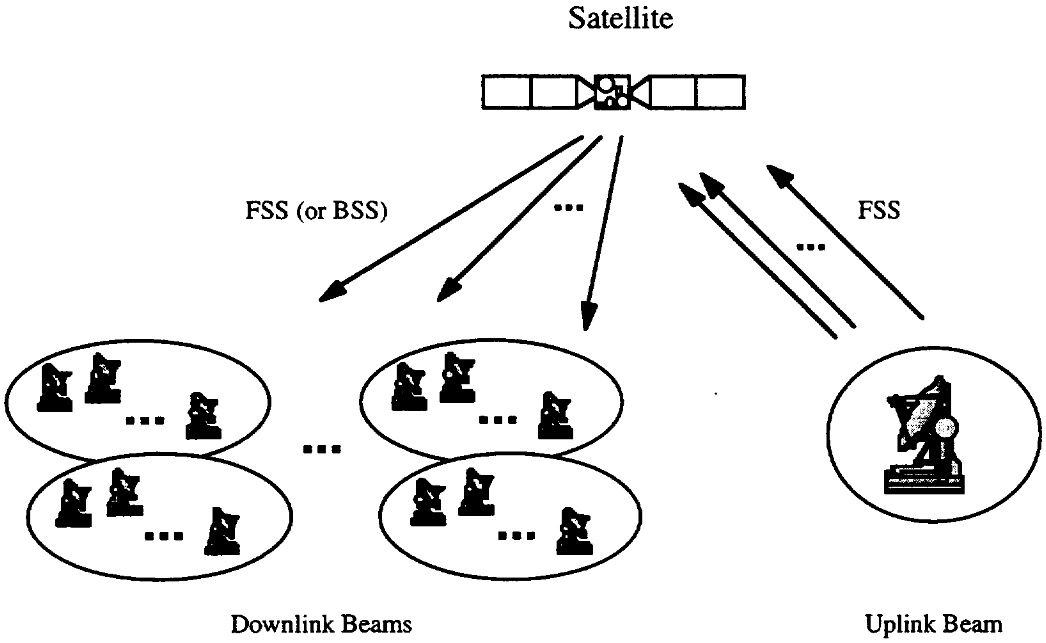


Figure 4.6 Satellite Video-On-Demand Service: Network Architecture VOD-D (For Video Delivery Links)

The signaling links are two-way links that are used for signaling purposes between the service provider and the subscribers. The architecture for the signaling links, Architecture VOD-S shown as Figure 4.7, has one downlink beam and one uplink beam. Through these links, the subscribers can check their account balance, view a list of video programs available, request selected video programs to be delivered to them at certain days and times, etc. The protocols used to establish these signaling links are very similar to those used in commercial networks (e.g., VSAT and Inmarsat). Typically, the outbound signaling links to the subscribers are TDM links where many subscribers receive the same carrier and extract the signaling information relevant to themselves with their own crypto keys. There are one or more inbound signaling carriers which the subscribers may access in a pre-assigned manner or contention manner in time (e.g., Aloha, Slotted Aloha and TDMA).

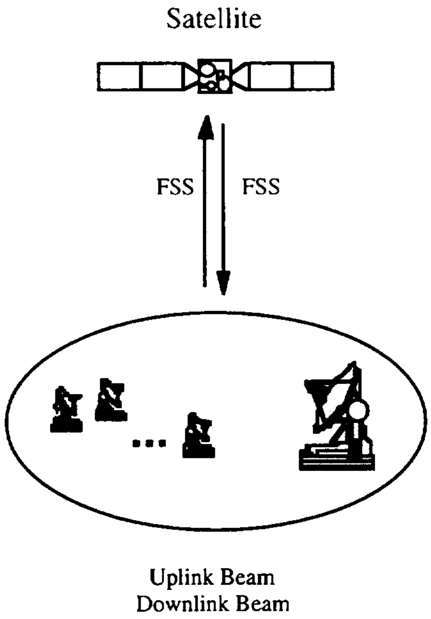


Figure 4.7 Satellite Video-On-Demand Service: Network Architecture VOD-S (For Signaling Links)

4.4.2 Satellite Payload Architectures

To support the video-on-demand service, four satellite payload architectures are addressed here to show the relevant technologies required: Architecture VOD1, Architecture VOD2, Architecture VOD3, and Architecture VOD4.

Architecture VOD1

The first architecture, Architecture VOD1 shown as Figures 4.8, is a "bent-pipe" architecture where no onboard processing is employed. In this architecture, a small portion of the satellite bandwidth is allocated to the signaling links. The remainder of the satellite bandwidth is divided into K sub-bands with each sub-band accommodating one or more video-delivery carriers. Depending on the requests, each of these sub-bands can be switched to any of the downlink beams, using a microwave switch matrix (MSM). Accordingly, the power required to support each beam varies with time. To optimize the satellite power required, active transmit phased array (ATPA) antennas are used. With ATPA antennas, the total available power from the satellite solid state power amplifiers (SSPAs) are automatically shared among the downlink beams to support their dynamic traffic (i.e., video programs). With ATPA antennas, the downlink beams can also be reconfigured and steered to desired locations on earth by adjusting the beam weights of the beam forming networks (BFNs) of the beam forming matrix (BFM). Due to the reconfiguration and steering capability of the beams, it may not be necessary to design the system with many beams.

MSM technology has been implemented at Ka-Band on the ACTS satellite [5.3-6] and other satellites (e.g., ITALSAT).

ATPA antennas have been flown on the Japanese ETS-6 satellite (which is an S-Band design) [4-29] and on Iridium satellites (which is an L-Band design). The ATPA

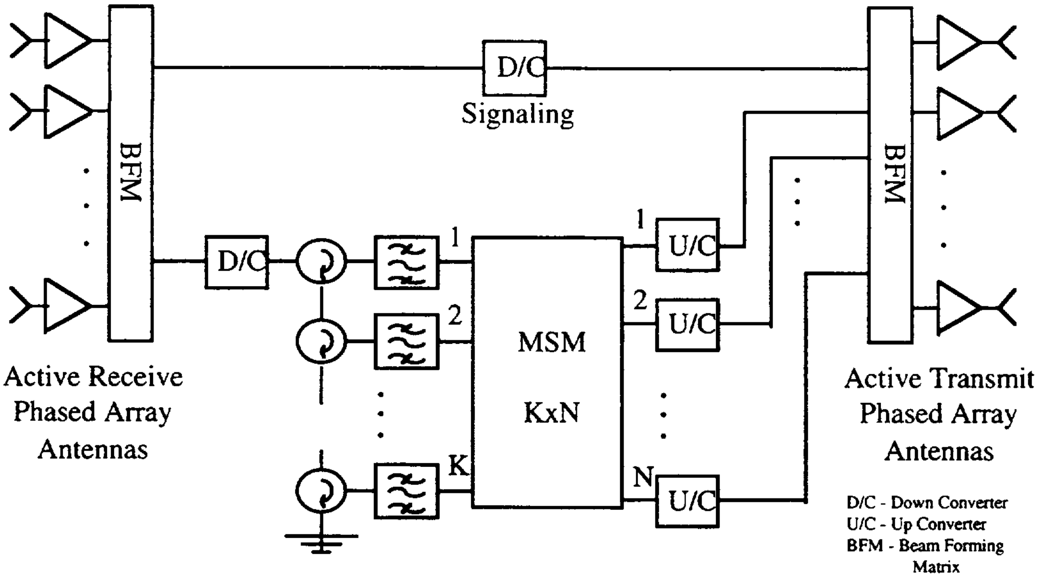

D/C
Signaling
BFM
BFM
D/C
1
2
K
MSM
KxN
1
2
N
U/C
U/C
U/C
Active Receive Phased Array Antennas
Active Transmit Phased Array Antennas
D/C - Down Converter
U/C - Up Converter
BFM - Beam Forming Matrix

Figure 4.8 Satellite Video-On-Demand Service: Satellite Payload Architecture VOD1

technology, is not mature, however, particularly in the area of implementation with monolithic microwave integrated circuits (MMIC) and photonics to reduce weight, and in the area of understanding the nonlinearity effects [4-30] to reduce power consumption.

Architecture VOD2

Architecture VOD2, shown as Figure 4.9, is very similar to Architecture VOD1. The differences are that:

- To share the RF power among the downlink beams, Architecture VOD2 uses MARS instead of ATPA antennas.
- Architecture VOD2 does not need to use MSM to do the switching because this switching capability can be replaced with MARS' routing capability.
- Architecture VOD2 uses reflector antennas to create downlink beams.

Architectures VOD3 and VOD4

Architectures VOD3 shown in Figure 4.10 and VOD4 shown in Figure 4.11 are the onboard regeneration versions of Architectures VOD1 and VOD2, respectively. As discussed earlier in Section 4.3, with onboard regeneration, the satellite DC prime power required is significantly reduced because:

- The uplink and downlink communication impairment are decoupled.
- Downlink rain fade can be compensated for with real-time downlink power control.

4.4.3 Supporting Technology Summary

The supporting technologies required for the satellite payload architectures for the satellite video-on-demand service are summarized in Table 4.13.

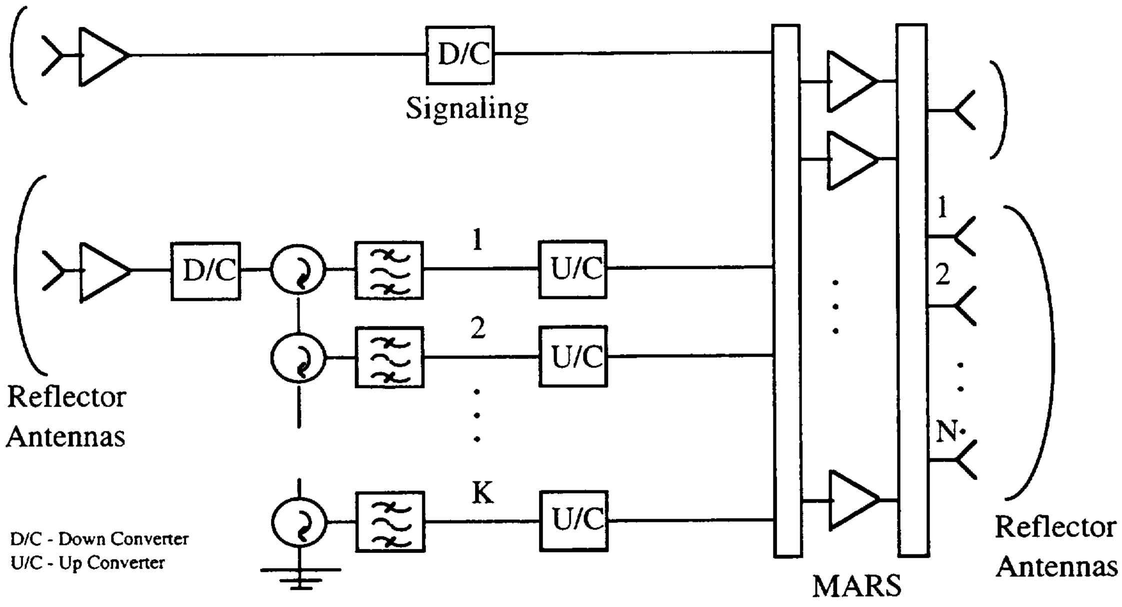


Figure 4.9 Satellite Video-On-Demand Service: Satellite Payload Architecture VOD2

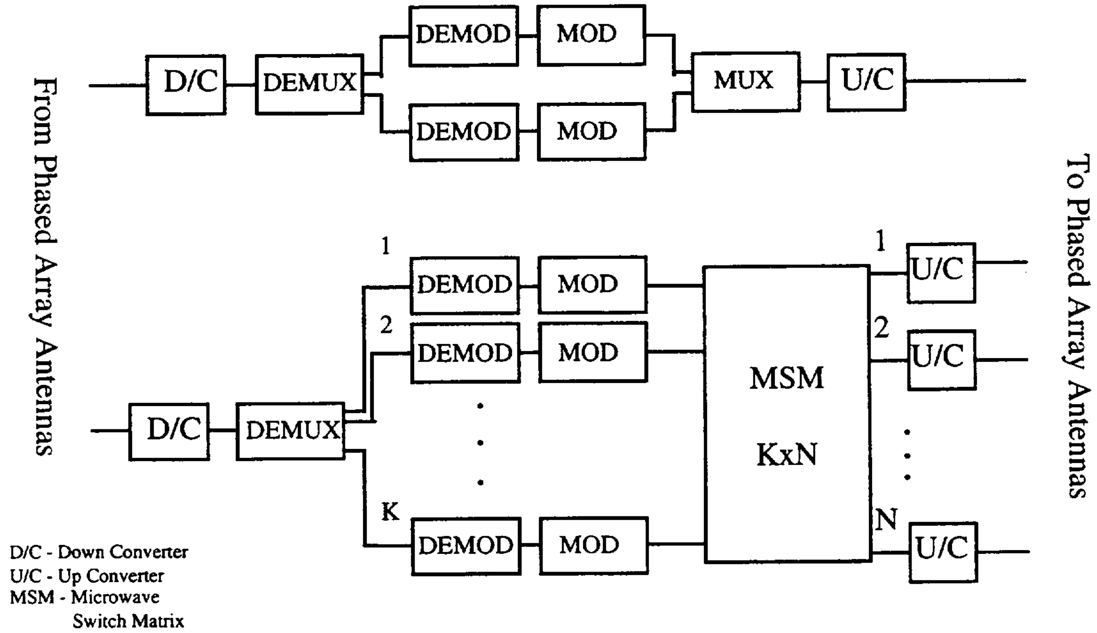


Figure 4.10 Satellite Video-On-Demand Service: Satellite Payload Architecture VOD3

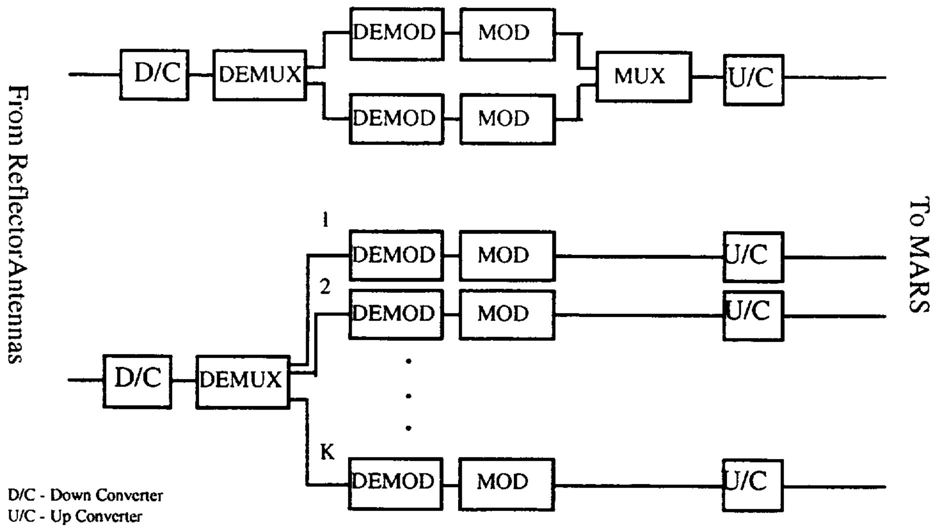


Figure 4.11 Satellite Video-On-Demand Service: Satellite Payload Architecture VOD4

Table 4.13
Supporting Technology for the Satellite Video-On-Demand Service

Satellite Payload Architecture	Supporting Technology	Comments
VOD1	• MSM • Active Transmit Phased Array (ATPA) (Including BFM) • Onboard Filter/Demultiplexer	• MSM - older technology but provides connectivity and service flexibility by user area- can use SAW technology to reduce weight and volume • ATPA requires development, particularly at Ka-band; could provide power /coverage flexibility for user services; also could provide adaptive capabilities to accommodate large dynamic fade range at Ka-band
VOD2	• MARS • Onboard Filter/Demultiplexer	• Same as VOD1 except MARS replaces ATPA-MSM. • Both ATPA and MARS are competing technology alternatives, particularly for Ka-band
VOD3	• MSM • ATPA (Including BFM) • Onboard Regeneration • Onboard Multicarrier Demultiplexer	• Same as VOD1 except onboard demod/remod used to reduce terminal and satellite transmit power • Reduces satellite prime power requirements
VOD4	• MARS • Onboard Regeneration • Onboard Multicarrier Demultiplexer	• Same as VOD2 except on board satellite demod/remod used to minimize terminal and satellite RF and prime power requirements
Common (VOD1, VOD2, VOD3, VOD4)	• Ka-band or 50/40 GHz Technology • Video Compression • Combined Video Compression/Video TDM • M-PSK and M-QAM Modulation with Turbo Codes	• Use of Ka-band or 50/40 GHz can be used with any of the above technologies, however, ATPA and MARS would require further development at these bands • Use of variable video compression and variable modulation/FEC rates and types would be useful at the higher frequency bands (e.g, 50/40 GHz)

4.5 SATELLITE BANDWIDTH-ON-DEMAND SERVICE

4.5.1 Satellite Network Architecture

Figure 4.12 is a satellite network architecture for the bandwidth-on-demand (BWOD) service described in Section 3. In this architecture, a user can demand satellite bandwidth and power for its communication to another user. The communication can be established with one satellite hop or with two satellite hops via two gateway stations. A user can also demand satellite bandwidth for connection to a public switched network, e.g., a public switched telephone network (PSTN), integrated service digital network (ISDN), or public land mobile network (PLMN).

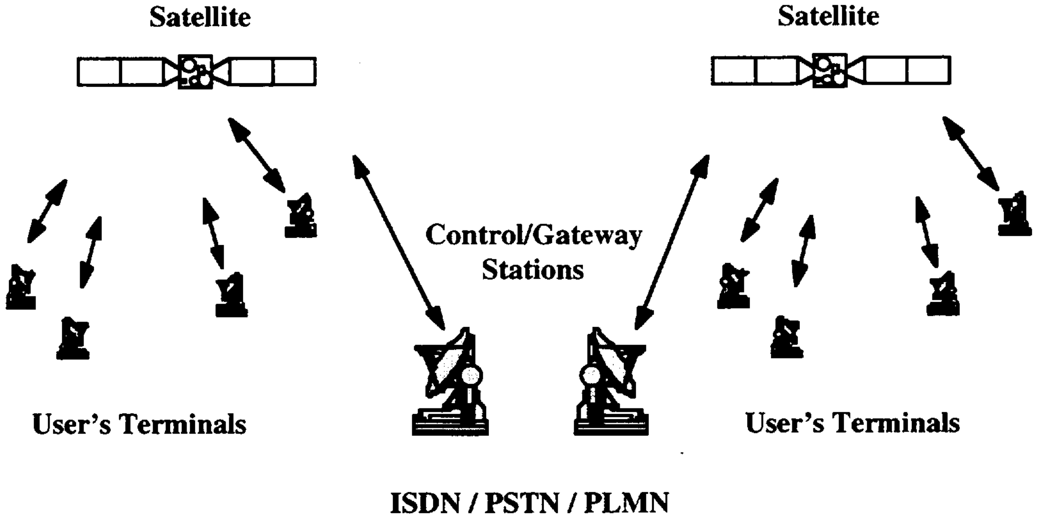


Figure 4.12 Satellite Bandwidth-On-Demand Service: Network Architecture BWOD

The satellite has M uplink beams which are fully connected with N downlink beams as shown in Figure 4.13.

4.5.2 Satellite Payload Architectures

To support the bandwidth-on-demand service, three satellite payload architectures are addressed here to show relevant technologies required: BWOD1, BWOD2 and BWOD3. The frequency plans for these three architectures are shown in Figure 4.14.

Architecture BWOD1

Architecture 1 uses the frequency plan shown in Figure 4.14. It divides the satellite available bandwidth into MxN sub-bands with a set of fixed bandwidths for full connection between uplink beams and downlink beams and a set of L sub-bands, each of which can be reassigned to connect any uplink and downlink beam in response to traffic demand. A user is assigned a bandwidth that can be a portion of either the fixed sub-bands or a reassignable sub-band. Through the fixed sub-bands, users and control stations can send signaling information.

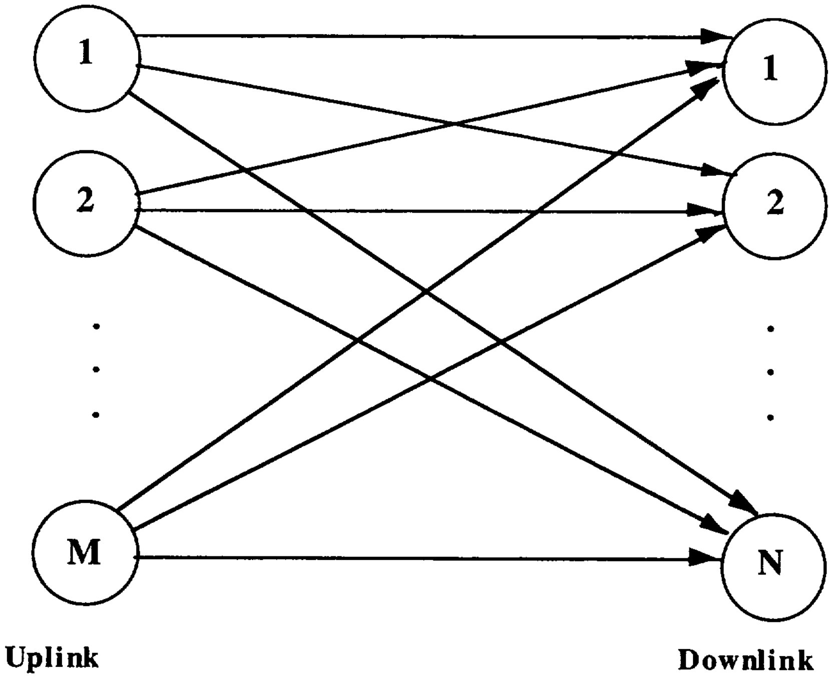
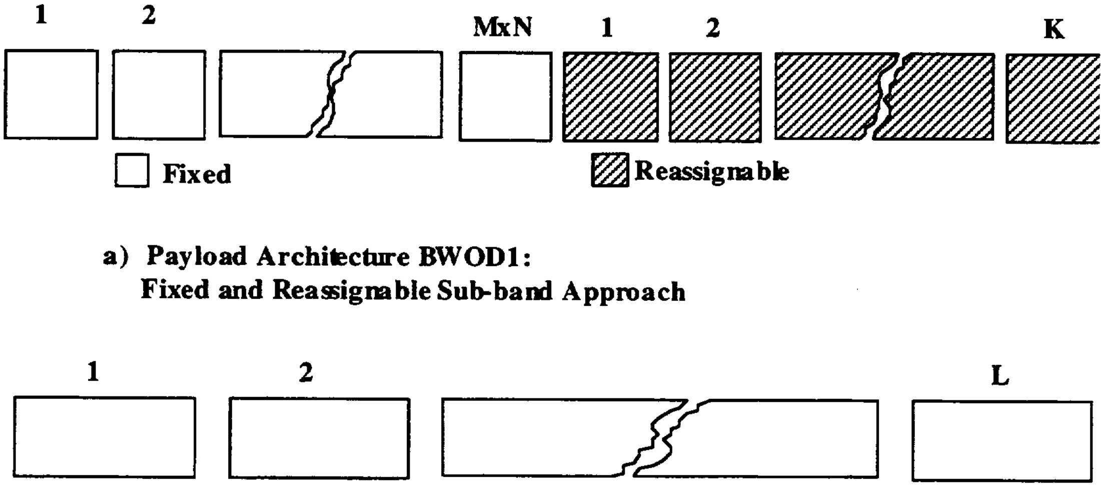


Figure 4.13 Satellite Bandwidth-On-Demand Service:
Satellite Antenna Beam Connection



a) Payload Architecture BWOD1:
Fixed and Reassignable Sub-band Approach

b) Payload Architecture BWOD2:
Multicarrier SS-TDMA (SS-TDMA/FDMA)

c) Payload Architecture BWOD3:
Multicarrier SS-TDMA with Onboard Regeneration Approach

Figure 4.14 Satellite Bandwidth-On-Demand Service:
Frequency Plans

There are two variations in Architecture BWOD1, namely Architecture BWOD1a (Figure 4.15a) and Architecture BWOD1b (Figure 4.15b). These two variations differ only in the antenna subsystems: Architecture BWOD1a utilizes active phased array antennas and Architecture BWOD1b utilizes reflector antennas with MARS to share the power among the downlink beams.

Architecture BWOD1 requires the use of one filter matrix for the fixed beam connection with fixed sub-bands and one switch-filter-switch matrix for the flexible beam connection with reassignable sub-bands. The filter matrix is an MxN matrix which has M (1-to-N) splitters, MxN filters, and N (M-to-1) combiners. The switch-filter-switch-filter matrix consists of an MxK MSM (microwave switch matrix discussed earlier), K filters, and KxN MSM connected in tandem.

Architecture BWOD2

Architecture BWOD2, shown in Figure 4.16, utilizes the SS-TDMA concept that was demonstrated with the NASA ACTS satellite [4-26] and other satellites (e.g., ITALSAT and INTELSAT). The architecture consists of L MSMs to correspond to L TDMA carriers. Corresponding to each TDMA carrier there are M identical filters for a total of MxL filters. While the MSMs in Architecture BWOD1 may change their switching states once or twice every day or every week, the MSMs for Architecture BWOD2 change their switching states every faction of a millisecond according to the TDMA burst time plans.

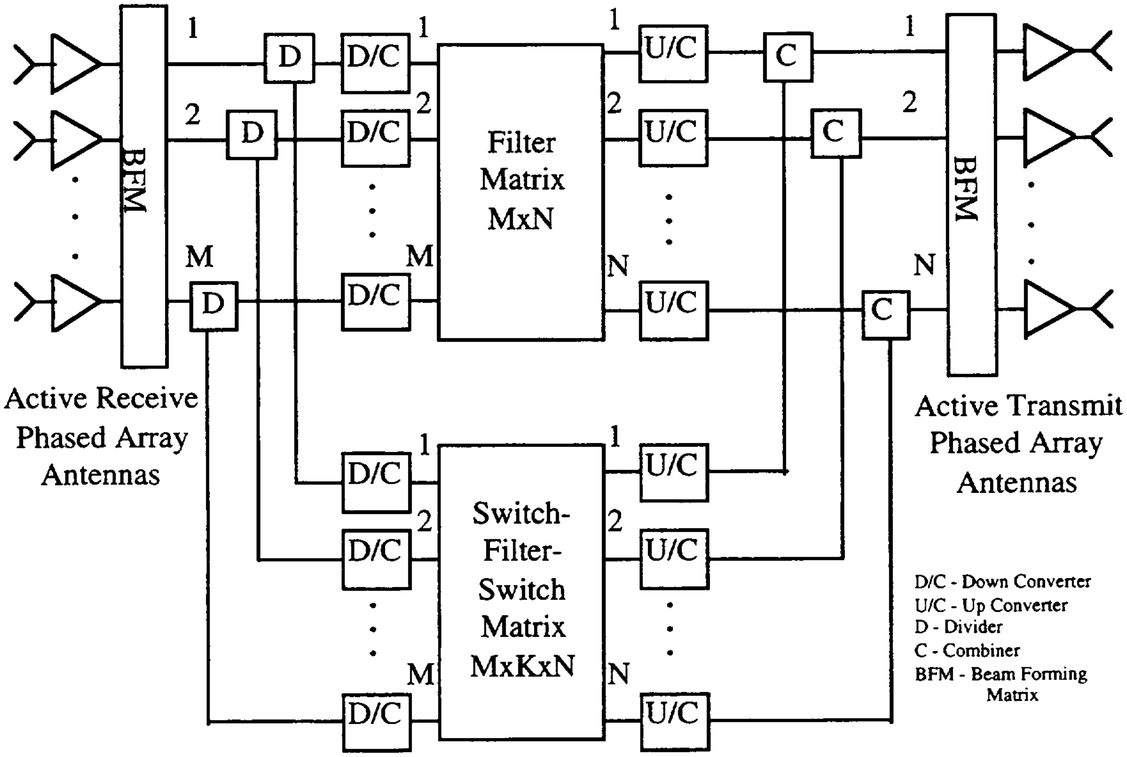


Figure 4.15a Satellite Bandwidth-On-Demand Service:
Satellite Payload Architecture BWOD1a with Active Phased Array Antennas

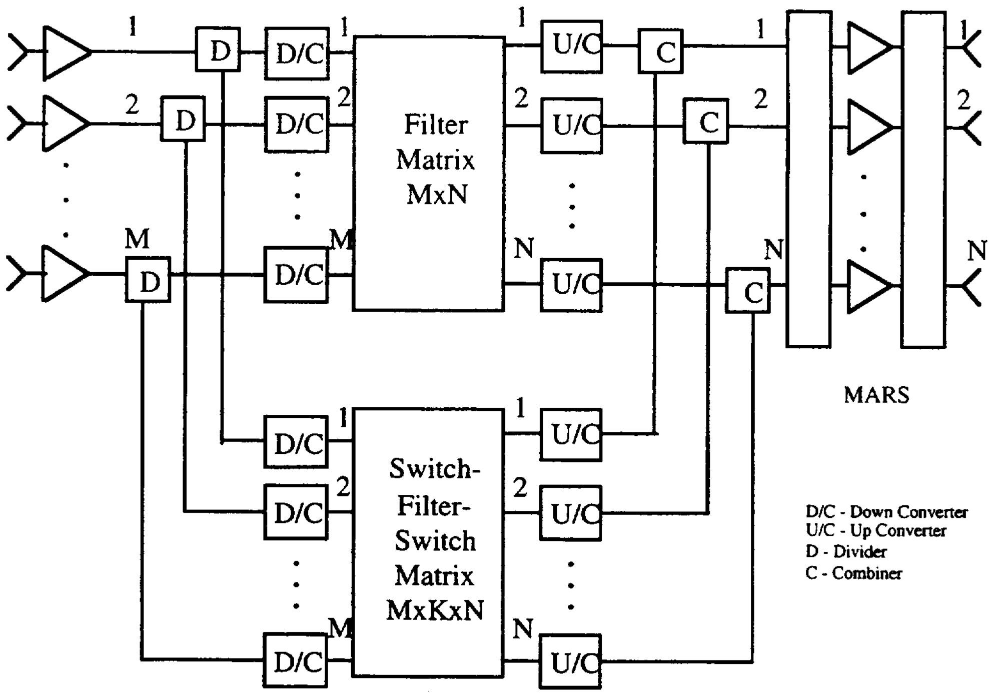


**Figure 4.15b Satellite Bandwidth-On-Demand Service:
Satellite Payload Architecture BWOD1b with Reflector Antennas and MARS**

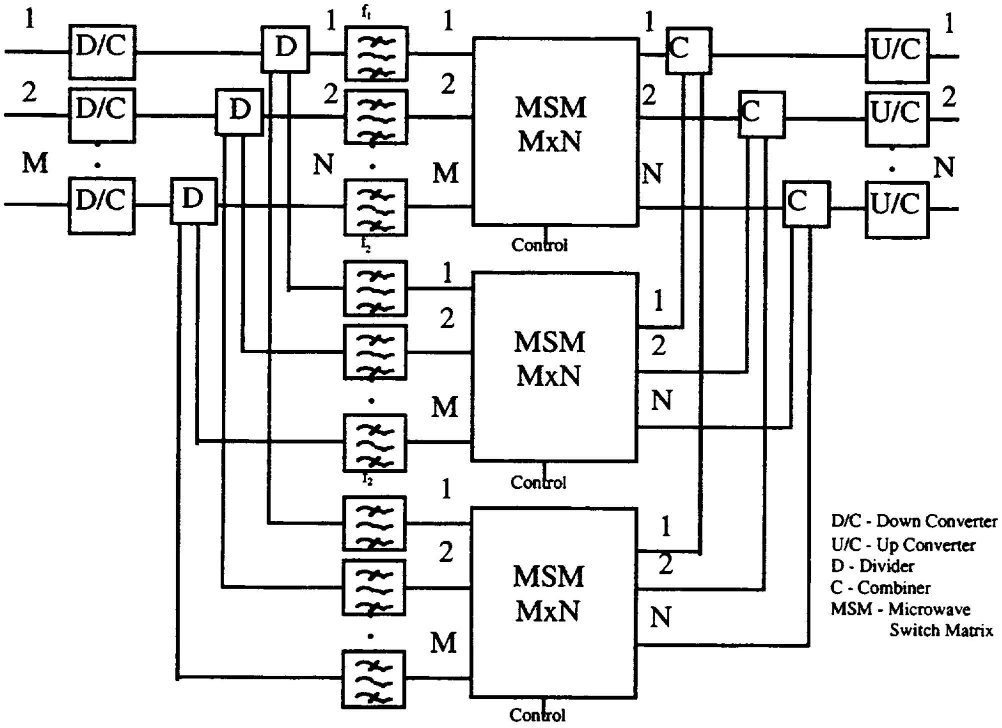


**Figure 4.16 Satellite Bandwidth-On-Demand Service:
Satellite Payload Architecture BWOD2 with Multicarrier SS-TDMA**

Architecture BWOD2 utilizes the bandwidth very efficiently. However, due to the TDMA nature, the earth station is required to burst data at a bit rate much higher than the information rate, and therefore Architecture BWOD2 requires the earth station to use a much larger antenna size and a much higher HPA power. The latter requirement can be mitigated by using variable burst rates, matched to the size of the terminal. This matching can be further enhanced by using variable modulation and FEC formats depending upon the terminal size, required information rate, and link fade depth status on each burst.

Architecture BWOD3

Architecture BWOD3 utilizes the combined SS-TDMA and onboard regeneration concept that was also demonstrated with the NASA ACTS satellite [4-26]. The satellite Architecture BWOD3 is very similar to Architecture BWOD2 with each MSM being replaced by a baseband switch matrix (BSM), M demodulators and N modulators as shown in Figure 4.17. Note, this is the same configuration as shown in Figure 4.16, with each MSM being replace by the following elements.

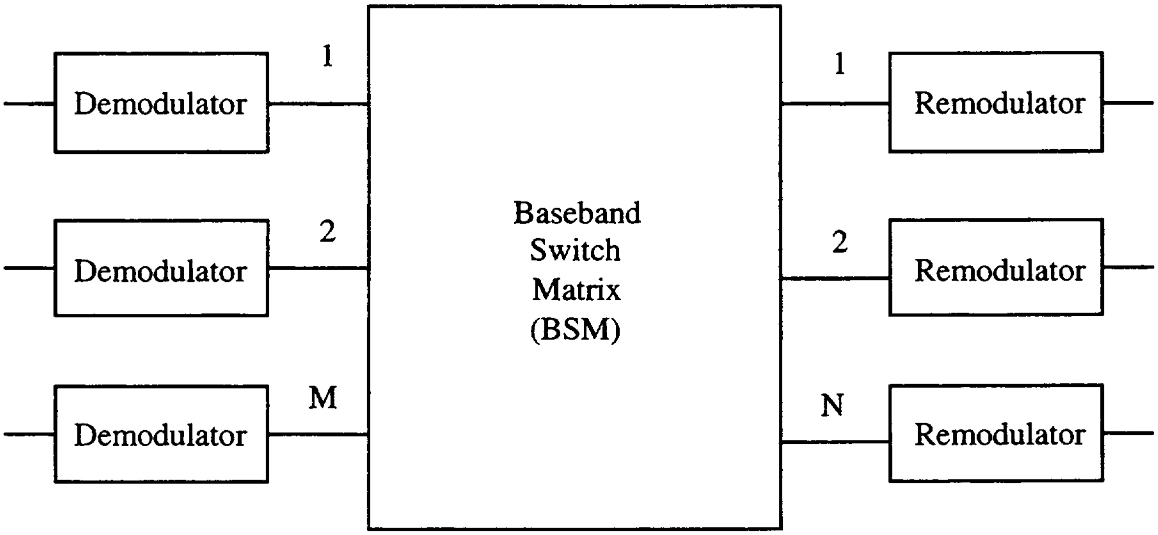


Figure 4.17 Satellite Bandwidth-On-Demand Service: Satellite Payload Architecture BWOD3 with Multicarrier SS-TDMA and Onboard Regeneration

4.5.3 Supporting Technology Summary

The supporting technologies required for the satellite payload architectures for the satellite bandwidth-on-demand service are summarized in Table 4.14.

Table 4.14
Supporting Technology for the Satellite Bandwidth-On-Demand Service

Satellite Payload Architecture	Supporting Technology	Comments
BWOD1a BWOD1b	• MSM • Onboard Filter/Demultiplexer • ATPA downlink • MARS replaces ATPA in BWOD1b	• Provides flexible analog assignment of power and bandwidth to varying user connectivity requirements- permits future ground-based changes in modulation and FEC • Baseline uses existing technology MSM that is re-configured daily or weekly, but could also rely on SAW filter & Switch matrix (SFSM)
BWOD2	• MSM (or SFSM) • SS-TDMA • Onboard Filter/Demultiplexer	• Same as BWOD1a and BWOD1b except that bandwidth and antenna beam connectivity switching occurs on a burst by burst basis vice on a weekly (or daily basis)
BWOD3	• BSM • SS-TDMA • Onboard Regeneration • Onboard Multicarrier Demultiplexing (FDM)	• Same as BWOD2 except use of onboard demodulation and re-modulation to reduce terminal and satellite transmit RF and prime power
Common (BWOD1, BWOD2, BWOD3)	• Ka-band or 50/40 GHz Technology • Active Phased Array (Including BFM) or MARS • M-PSK and M-QAM Modulation with Turbo Codes	• All architectures can take advantage of Ka and 50/40 GHz bands, however, ATPA and MARS technologies require development • Use of adaptive modulation and FEC can be achieved with the analog switched bandwidth concepts in (BWOD1a, BWOD1b, and BWOD2, but technology development would be required for satellite adaptive multi-modulation and FEC in BWOD3

5. SUMMARY, CONCLUSIONS AND RECOMMENDATIONS

In this report, services that may be offered by the generation-after-next SATCOM systems were identified and described. These services include TV broadcast (TVB), video-on-demand (VOD), and bandwidth-on-demand (BWOD). Different top-level satellite payload architectures for these services were generated to show the relevant technologies required. The conclusions and recommendations that can be drawn from this report regarding technologies that may be used to support the generation-after-next SATCOM systems are presented below.

a) Because the C-band and Ku-band have been fully utilized, the generation-after-next SATCOM systems are likely to be operating at Ka-band, 50/40 GHz band or at optical bands. Technologies associated with Ka-band have been demonstrated with the NASA ACTS satellite and other satellites (e.g., Italsat). The Ka-band technologies will be mature when the next-generation Ka-band multimedia satellite systems are deployed (e.g., Celestri, Spaceway/Galaxy and Teledesic).

Technologies associated with the 50/40 GHz band are relatively new. They are practically limited to propagation experiments being conducted with the Italsat satellite.

Laser communication technologies have been investigated for space use by NASA and other organizations. Other than the availability of large bandwidth, space laser communication technologies, as compared to microwave technologies, may offer a 3:1 advantage in weight, 2:1 in required power, and immunity to interference, jamming, and interception. Earth-to-space and intersatellite laser communication experiments were planned on Japan's ETS-6 satellite [4-9]. Due to failure of the launch of the ETS-6 satellite in 1996, however, those experiments were not conducted. Laser communications are greatly affected by the earth's atmosphere, which causes the laser beam to wander. Accordingly, laser communications can be used for intersatellite links, but for earth-space communications, more studies are required to fully understand the atmospheric effects and techniques that can be used to compensate for these effects.

Therefore, it is recommended that:

- The 50/40 GHz technologies for different transmission components (e.g., HPA, LNA, filter, mixer), fabrication techniques (MMIC, MIC, microstrip, TEM-line), and earth-space propagation be further studied to extend them to full maturity.
- The laser technologies for different transmission components and earth-space propagation be further studied to extend them to full maturity. The earth-space propagation experiments can be performed with a laser beacon placed onboard a space platform (e.g., a GEO satellite, LEO satellite, or a space station).

b) For the TVB and VOD services, there are no technology challenges other than those (to be addressed in d and e below) that make more efficient use of the available satellite bandwidth and power. Although each of these services may have high bit-rate requirements[4] which may not be met by the use of a single GEO satellite[5], the requirements can be decomposed to requirements for different geographical areas each of which can be served by a different GEO satellites. This decomposition of requirements can satisfy TVB and VOD services because these services only require "hub-to-spoke" communication from one or more distribution centers to users' locations.

c) For the BWOD service, there are challenges in technologies other than those that make more efficient use of the available satellite bandwidth and power. These challenges arise because BWOD service has: i) extremely high bit-rate requirements[6]; ii) flexible connection requirements between any two locations in the service area; and, iii) low time-delay requirements in certain types of the BWOD service (e.g., interactive

[4] Estimated from Section 3 to be 2 and 2.4 Gbps respectively for the TVB and VOD services in the US.

[5] The bit-rate capacity of a satellite depends on how a satellite is designed, the amount of bandwidth allocated to the satellite (by the ITU), and the constraints on power, weight and size by available satellite buses.

[6] Estimated from Section 3 to be 10 Gbps for a US metropolitan area.

data communication). To meet the requirements of i and ii, use of many satellites (GEOs and/or LEOs) is necessary. To meet the requirements of iii, direct links from these satellites (i.e., intersatellite links or crosslinks), instead of indirect links via gateway stations, are likely necessary. It is recommended that:

- The GEO satellite payload architectures derived for the BWOD service in Section 4 be enhanced to accommodate intersatellite links.
- LEO satellite constellations and their associated payload architectures be generated for the BWOD service. These architectures may be based on those designed for the next generation Ka-band satellite systems (e.g., Celestri or Teledesic) with enhancement using some of the technologies mentioned in d and e below.

d) To meet the high bit-rate requirements for the TVB, VOD and BWOD services, it is necessary to use have a large amount of bandwidth and to use this bandwidth efficiently.

To have a large amount of bandwidth, it is necessary to reuse the frequencies allocated by the ITU as many times as is economically possible and also make use of higher order modulation and high code rate FEC technologies. Frequencies can be reused in three different ways: i) through antenna polarization discrimination where the same frequencies may be used by the same satellite to serve the same geographical area; ii) through spatial satellite antenna discrimination where the same frequencies can be used by the same satellite on different geographical areas; and, iii) through earth station antenna off-axis gain discrimination where different satellites can use the same frequencies to serve the same geographical area. The number of times that the frequencies can be reused depends on the beamwidths of the satellite antenna beams and on discrimination levels, which depends on the designs of the earth station and satellite antennas. The amount of frequency reuse also depends on the amount of interference that the communication links can tolerate, which in turn depends on the modulation and coding used and the transmission quality required. Although substantial benefits will accrue through frequency reuse, the design of satellite payloads becomes more complicated due to the need for more filtering, switching and routing that must be performed. It is recommended that:

- A study be performed to determine under what operating conditions, if any, that a LEO system can share the same frequencies (reuse the same frequencies) with another LEO system or a GEO system.
- A study be performed to determine the realistic number of times that the same frequencies can be reused for the GEO systems that operate at Ka-band and 50/40 GHz band to support the TVB, VOD and BWOD services.
- The satellite payload architectures generated in Section 4 be enhanced to take frequency reuse into consideration.
- To use the available satellite bandwidth efficiently, it is necessary to use a bandwidth efficient modulation such as 8-PSK, 16-QAM, 32-QAM and even higher order modulation formats up to 256-QAM. Use of these modulation formats can significantly increase the amount of required satellite and earth station power, as compared to the standard modulations, such as BPSK and QPSK used in satellite communications. Thus, appropriate FEC coding must

be used in conjunction with these modulation formats to balance the bandwidth and power requirements. Therefore, it is recommended that a tradeoff study be performed to determine the best combinations of modulation (e.g., QPSK, 8-PSK, 16-QAM, and 256-QAM) and FEC techniques with different codes (e.g., convolutional, concatenated {convolutional and block codes}, and Turbo) and different code rates (e.g., 1/2, 2/3, 3/4, and 7/8, for convolutional and Turbo FEC) with high rate block codes (e.g., Reed-Solomon 233/255), and variable length interleavers to support a given bit-rate with a given available satellite bandwidth. The use of dynamic assignment of modulation and FEC (code and interleaver assignments) can be used to compensate for the potentially large link attenuations that could be experienced at higher frequency bands (Ka and 50/40 GHz). This dynamic assignment can also be made consistent with planned changes to ATM standards that will achieve greater efficiency through use of the variable bit rate portions of these standards.

Another way to efficiently use the available satellite bandwidth is to dynamically allocate the bandwidth among the uplink beams and downlink beams to reflect changes in communication traffic load. That is, it is desirable to design a satellite payload with a bank of filters whose center frequencies and bandwidths are adjustable and with microwave or baseband switch matrices and MARS that are used to connect these filters to appropriate uplink and downlink beams. It is recommended that:

- A tradeoff study be performed among the bulk digital filtering techniques (e.g., polyphase, and FFT/IFFT) used for frequency division multiplexing with respect to speed, flexibility, frequency response.
- A/D and D/A technologies be further researched to improve sampling speed.
- Microwave and baseband switch matrices be further researched to improve switching speed.
- The MARS technology be further researched to fully understand its routing and power distribution capability.

e) To meet the high bit-rate requirements for the TVB, VOD and BWOD services, it is necessary to have a large amount of satellite power and to use this power efficiently.

To have a large amount of satellite power, it is necessary to have a large solar array with improved efficiency in solar-to-electric power conversion. To efficiently use the satellite power, it is necessary to: i) employ a power-efficient modulation/coding technique (e.g., Turbo code); ii) use active transmit phased array (ATPA) antennas or MARS so that the available RF power available among the downlink beams are automatically shared; iii) develop techniques that can estimate the transponder nonlinearity effects accurately so that the HPAs can be operated as close to their saturation levels as possible; iv) use of onboard regeneration; v) use of uplink and downlink power control to compensate for rainfade for Ka-band and 50/40 GHz operation; vi) use of narrow spot beams; and, vii) use of lower power electronics. Thus, it is recommended that:

- Further research on Turbo codes be continued for full understanding of its capability and for implementation into integrated circuits.
- Tradeoff studies between the two existing satellite power sharing techniques,

ATPA and MARS, be conducted.

- Software to analyze the nonlinearity/intermod effects of an ATPA and a MARS [4-14, 4-16] be funded for enhancement.
- Investigation of effective means to relay downlink rainfall effects to satellite be conducted.
- Further research on improvement of narrow beam pointing accuracy be conducted.

References

[3-1] Mid-Term Report, Assessment of Broadband Satellite markets and Technologies, NASA Lewis Research Center, July, 1997, p. A-1

[3-2] Worthman, Ernest, "Direct Broadcast Satellites Find Niche in Consumer Electronics Market", RF Design Magazine, August 1997, pp. 36-40.

[3-3] Rath, Kamlesh, D.H. Wanigasekara, R.G. Wendorf, D.C. Verma, "Interactive Digital Video Networks: Lessons from a Commercial Deployment", IEEE Communications Magazine, June 1997, pp. 70-74.

[3-4] Kwok, Timothy C., "Residential Broadband Internet Services and Applications Requirements", IEEE Communications Magazine, June 1997, pp. 76-83.

[4-1] S. Sharrock, "View From Europe: Shall We Share?" Satellite Communications, June 1997.

[4-2] Celestri Multimedia LEO System, FCC Filing, June 1997.

[4-3] Radio Regulations, International Telecommunications Union, Geneva, 1996.

[4-4] Tables of Frequency Allocations and Other Extracts From: Manual of Regulations and Procedures for Federal Radio Frequency Management. U.S. Department of Commerce, National Telecommunications and Information Administration (NTIA), September 1995 Edition.

[4-5] J. Careless, "Ka-band Satellites - A Guide to the Proposed U.S. Systems," Via Satellite, February 1996.

[4-6] "M-Star," www.nfra.nl/craf/mstar.html.

[4-7] "V-Stream," PanAmSat News Release, Florida Space Today, Oct. 2, 1997 Issue.

[4-8] G. Hyde and B.I. Edelson, "Laser SATCOM Offers Radio Links in Space," Aerospace America, September 1997.

[4-9] K. Araki, Y. Arimoto, M. Shikantani, M. Toyoda, and T. Aruga, "ETS-VI Laser Communications Experiment System," Proceedings of the 14th AIAA International Communication Satellite Systems Conference and Exhibit, Washington, D.C., March 22 - 26, 1992.

[4-10] X.T. Vuong and S.T. Vuong, "Satellite Link Margin and Availability Issues," IEEE Transactions on Broadcasting, June 1997.

[4-11] L.J. Ippolito, R.D. Kaul and R.G. Wallace, Propagation Effects Handbook for Satellite Systems Design, NASA Publication 1082.

[4-12] C. Berrou et al., "Near Shannon Limit Error Correcting Coding and Decoding: Turbo Codes (1)," Proceedings of IEEE International Conference on Communications, June 1993.

[4-13] D. Divsalar, "Turbo Codes," Tutorial Material Presented at the 1996 IEEE Military Communications Conference, October 1996.

[4-14] X.T. Vuong, "Active Transmit Phase Array Analysis Program (ATPAL)," Proceedings of the IEEE International Symposium on Phased Array Systems and Technology, Boston, MA, October 15 - 18, 1996.

[4-15] Matrix Amplifier and Routing System (MARS), Final Report prepared by SAIC for USAF Space and Missile Systems Center (SMC), Contract No. F25606-91-D0005, August 11, 1993.

[4-16] X.T. Vuong, "Matrix Amplifier and Routing System (MARS) Analysis Program," Proceedings of the IEEE Military Communications Conference (Milcom'97), Monterey, California, November 2 - 5, 1997.

[4-17] Y.S. Lee, A.E. Atia, and D.S. Ponchak, "Intersatellite Link Application to Commercial Communications Satellites," COMSAT Technical Review, Fall 1988.

[4-18] Digital Broadcasting Systems for Television, Sound and Data Services; Framing Structure, Channel Coding and Modulation for 11/12 GHz Satellite Services, European Telecommunication Standard ETS 300 421, European Telecommunications Standards Institute (ETSI).

[4-19] D.J. Whalen and G. Churan, "The American Mobile Satellite Corporation Space Segment," Proceedings of the 16th AIAA International Communication Satellite Systems Conference and Exhibit, Washington, D.C., Feb. 25 - 29, 1992.

[4-20] C. Hoeber, J. LaPrade and J. Morris, "Project Omega: A Case Study in Spacecraft Product Improvement," Proceedings of the 16th AIAA International Communication Satellite Systems Conference and Exhibit, Washington, D.C., Feb. 25 - 29, 1992.

[4-21] D.L. Wright and J.R. Balombin, "ACTS System Capability and Performance," Proceedings of the 14th AIAA International Communication Satellite Systems Conference and Exhibit, Washington, D.C., March 22 - 26, 1992.

[4-22] R.T. Gedney et al., "Operational Uses of ACTS Technology," Proceedings of the 14th AIAA International Communication Satellite Systems Conference and Exhibit, Washington, D.C., March 22 - 26, 1992.

[4-23] M. Mauri, B. Giannone and A. Paraboni, "Preliminary Results of the ITALSAT Propagation Experiment in 20, 40 and 50 GHz Bands," Proceedings of the 14th AIAA International Communication Satellite Systems Conference and Exhibit, Washington, D.C., March 22 - 26, 1992.

[4-24] D.H. Layer, J.M. Kappes, and C.B. Cotner, "140-Mbit/s COPSK Modem Laboratory Tests and Transatlantic Field Trials," COMSAT Technical Review, Number 1, Spring 1990.

[4-25] Bandwidth Efficient Modulation Modem Test Plan, Draft Report prepared by SAIC for Defense Information Systems Agency (DISA), Contract No. DCA100-94-C-0079, April 19, 1996.

[4-26] W.F. Cashman, "ACTS Multibeam Communications Package: Description and Performance Characterization," Proceedings of the 14th AIAA International

Communication Satellite Systems Conference and Exhibit, Washington, D.C., March 22 - 26, 1992.

[4-27] P. Cangiane, H.A. Courtois, D.D. Lee, M.A. Sherry, and M.E. Spencer, "A Multi-Channel Demultiplexer/Demodulator Architecture, Simulation and Implementation," Proceedings of the 14th AIAA International Communication Satellite Systems Conference and Exhibit, Washington, D.C., March 22 - 26, 1992.

[4-28] G.R. Welti, "Butler Matrix Transponder Improvements," US Patent Application, Serial No. 412399, Nov. 1972.

[4-29] Masato et al., "Antenna Pattern Measurement of the S-Band Active Phased Array on the ETS-VI," Proceedings of the 16th AIAA International Communication Satellite Systems Conference and Exhibit, Washington, D.C., Feb. 25 -29, 1996.

[4-30] X.T. Vuong, "Active Transmit Phase Array Analysis Program (ATPAL)," Proceedings of the IEEE International Symposium on Phased Array Systems and Technology, Boston, MA, October 15 - 18, 1996.

[5-1] G. Hyde and B.I. Edelson, "Laser SATCOM Offers Radio Links in Space," Aerospace America, September 1997.

[5-2] K. Araki, Y. Arimoto, M. Shikantani, M. Toyoda, and T. Aruga, "ETS-VI Laser Communications Experiment System," Proceedings of the 14th AIAA International Communication Satellite Systems Conference and Exhibit, Washington, D.C., March 22 - 26, 1992.

[5-3] X.T. Vuong, "Active Transmit Phased Array Analysis Program (ATPAL)," Proceedings of the IEEE International Symposium on Phased Array Systems and Technology, Boston, Massachsetts, Oct. 15 - 18, 1996.

[5-4] X.T. Vuong, "Matrix Amplifier and Routing System (MARS) Analysis Program," Proceedings of the IEEE Military Communications Conference, Monterey, California, Nov. 2 - 5, 1997.

ANNEX 1. ACRONYMS AND ABBREVIATIONS

ACTS	Advanced Communications Technology Satellite
ADPCM	Adaptive Differential PCM
AFM	Atomic Force Microscope
AI	Artificial Intelligence
AM	Amplitude Modulation
ARPA	Advanced Research Projects Agency
ASIC	Application-specific Integrated Circuit
ASIM	Application-specific Integrated Microinstrument
ATM	Asynchronous Transfer Mode
ATPA	Active Transmit Phased Array
AWGN	Additive White Gaussian Noise
BESOI	Bond and Etch-back of SOI
BFM	Beam Forming Matrix
BFN	Beam Forming Network
BOD	Bandwidth-on-demand
BPSK	Binary Phase Shift Keying
BSM	Baseband Switch Matrix
BSS	Broadcast Satellite Services
BTLZ	V.42bis Modem Compression Algorithm
CAD	Computer-aided Design
CD-ROM	Compact disk, Read-only Memory
CELP	Code-excited Linear Prediction
CMOS	Complementary Metal Oxide Semiconductor
CVSD	Continuously-variable Slope Delta Encoding
CARPA	Defense Advanced Research Projects Agency
D/C	Down-Converter
DC	Direct Current
DCT	Discrete Cosine Transform
DEMOD	Demodulation
DEMUX	Demultiplexer

DISA	Defense Information Systems Agency
DNA	Deoxyribonucleic Acid
DoD	Department of Defense
DRAM	Dynamic Random Access Memory
DSCS	Defense Satellite Communication System
DVB-S	Digital Video Broadcast - Satellite Format
EESS	Earth Exploration Satellite Services
EIRP	Effective Isotropically Radiated Power
FDM	Frequency Division Multiplex
FDMA	Frequency Division Multiple Access
FEC	Forward Error Correction
FFT	Fast Fourier Transform
FM	Frequency Modulation
FS	Fixed (Terrestrial) Services
FSS	Fixed Satellite Services
GEO	Geostationary Earth Orbit
GIF	Graphics Interchange Format
GSO	Geostationary Satellite Orbit
HARC-C	Houston Advanced Research Center Compression Algorithm
HDTV	High Definition Television
HEO	Highly Elliptical Earth Orbit
HPA	High Power Amplifier
IEEE	Institute of Electrical and Electronics Engineers
IFFT	Inverse Fast Fourier Transform
ISDN	Integrated Service Digital Network
ISO	International Standards Organization
ITU	International Telecommunications Union
JPEG	Joint Photographic Experts Group
JPL	Jet Propulsion Laboratory
LCD	Liquid Crystal Display
LNA	Low Noise Amplifier
LPC	Linear Predictive Coding

LPE	Low Power Electronics
LZW	Lempel, Ziv, Welch Compression Algorithm
MARS	Matrix Amplifier and Routing System
MBE	Molecular Beam Epitaxy
MCPC	Multiple Channels Per Carrier
MELP	Mixed Excitation Linear Predictive Vocoder
MEMS	Micro Electro-mechanical Systems
MEO	Medium Earth Orbit
MITI	Ministry of International Trade and Industry (Japan)
MMIC	Monolithic Microwave Integrated Circuit
MNP	Microcom Networking Protocol
MOD	Modulator
MOS	Metal Oxide Semiconductor
MOSFET	Metal Oxide Semiconductor Field-effect Transistor
MPEG	Motion Picture Experts Group
MS	Mobile (Terrestrial) Services
MSM	Microwave Switch Matrix
MSS	Mobile Satellite Services
MUX	Multiplexer
NASA	National Aeronautics and Space Administration
NGSO	Non-GSO
NSF	National Science
NTSC	National Television System Committee
NVOD	Near Video-on-demand
8-PSK	Octal Phase Shift Keying
PAL	Phase Alternating Line
PATMOS	Power and Timing Modeling Optimization and Simulation
PCA	Principal Component Analysis
PCM	Pulse-code Modulation
PDA	Personal Digital Assistant
PLMN	Public Land Mobile Network
PSTN	Public Switched Telephone Network

QAM	Quadrature Amplitude Modulation
QPSK	Quadrature (or Quaternary) Phase Shift Keying
RAM	Randon Access Memory
RAS	Radio Astronomy Services
RF	Radio Frequency
RNS	Radio Navigation Services
RNSS	Radio Navigation Satellite Services
RTG	Radio-isotope Thermoelectric Generator
SATCOM	Satellite Communication
SAW	Surface Acoustic Wave
SCPC	Single (Baseband) Channel Per Carrier
SECAM	Sequential Color with Memory (French acronym)
SIMOX	Separation by Implantation of Oxygen
SPICE	Simulation Program with Integrated Circuit Emphasis
SRAM	Static Randon Access Memory
SRS	Space Research Services
SSPA	Solid State Power Amplifier
SS-TDMA	Satellite Switched TDMA
STM	Scanning Tunneling Microscope
TDM	Time Division Multiplex
TDMA	Time Division Multiple Access
TV	Television
TVB	Television Broadcast
U/C	Up-Converter
ULTRA	Ultra Defense, Ultra Fast Computing Components Program (ARPA)
VOD	Video-on-demand
VSAT	Very Small Aperture Terminal

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY (*Leave blank*)

2. REPORT DATE
May 1998

3. REPORT TYPE AND DATES COVERED
Final Contractor Report

4. TITLE AND SUBTITLE

Technology Directions for the 21st Century
Volume IV

5. FUNDING NUMBERS

WU-315-90-81-00
NAS3-26565

6. AUTHOR(S)

Giles Crimi, Henry Verheggen, Robert Botta, Heywood Paul, and Xuyen Vuong

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Science Applications International Corporation
McLean, Virginia

8. PERFORMING ORGANIZATION REPORT NUMBER

E-11172

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

National Aeronautics and Space Administration
Lewis Research Center
Cleveland, Ohio 44135-3191

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

NASA CR—1998-207408

11. SUPPLEMENTARY NOTES

Project Manager, Denise S. Ponchak, Space Communications Office, NASA Lewis Research Center, organization code 6150, (216) 433-3465.

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Unclassified - Unlimited
Subject Categories: 17, 32, and 33 Distribution: Nonstandard

This publication is available from the NASA Center for AeroSpace Information, (301) 621-0390.

12b. DISTRIBUTION CODE

13. ABSTRACT ***(Maximum 200 words)***

Data compression is an important tool for reducing the bandwidth of communications systems, and thus for reducing the size, weight, and power of spacecraft systems. For data requiring lossless transmissions, including most science data from spacecraft sensors, small compression factors of two to three may be expected. Little improvement can be expected over time. For data that is suitable for lossy compression, such as video data streams, much higher compression factors can be expected, such as 100 or more. More progress can be expected in this branch of the field, since there is more hidden redundancy and many more ways to exploit that redundancy.

14. SUBJECT TERMS

Data compression; Low power electronics; Nanotechnology; Satellite communications

15. NUMBER OF PAGES
117

16. PRICE CODE
A06

17. SECURITY CLASSIFICATION OF REPORT
Unclassified

18. SECURITY CLASSIFICATION OF THIS PAGE
Unclassified

19. SECURITY CLASSIFICATION OF ABSTRACT
Unclassified

20. LIMITATION OF ABSTRACT